




3 1761 10374364 7



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743647>

12-001

Government
Publications

88

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

June 2009

•

Volume 35

•

Number 1



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and select "Publications."

This product, Catalogue no. 12-001-X, is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "Providing services to Canadians."

Survey Methodology

A journal
published by
Statistics Canada

June 2009 • Volume 35 • Number 1

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2009

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 2009

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Past Chairmen G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Members J. Gambino
R. Jones
J. Kovar
H. Mantel
E. Rancourt

EDITORIAL BOARD

Editor J. Kovar, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor M.P. Singh (1975-2005)

Associate Editors

J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
D. Judkins, *Westat Inc.*
D. Kasprzyk, *Mathematica Policy Research*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
S.M. Miller, *Bureau of Labor Statistics*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*
D. Pfeiffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*

J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

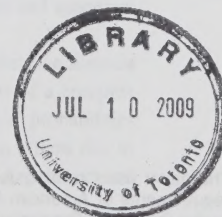
Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology
A Journal Published by Statistics Canada
Volume 35, Number 1, June 2009

Contents

In This Issue.....	1
 Regular Papers	
Victor M. Estevao and Carl-Erik Särndal A new face on two-phase sampling with calibration estimators	3
Jianzhu Li and Richard Valliant Survey weighted hat matrix and leverages	15
Jean-François Beaumont and Cynthia Bocci A practical bootstrap method for testing hypotheses from survey data.....	25
Bo-Seung Choi, Jai Won Choi and Yousung Park Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye State) Polls	37
Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff and Van L. Parsons Hierarchical and empirical Bayes small domain estimation of the proportion of persons without health insurance for minority subpopulations.....	53
Tucker McElroy and Scott Holan A nonparametric test for residual seasonality.....	67
Siegfried Gabler and Partha Lahiri On the definition and interpretation of interviewer variability for a complex sampling design	85
Barry Schouten, Fannie Cobben and Jelke Bethlehem Indicators for the representativeness of survey response.....	101
Guillaume Chauvet Stratified balanced sampling.....	115



to the fact that the paper used in this publication meets the minimum requirements of American National Standard for Information Sciences - Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.

Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences - "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.



In this issue

In the first paper of this issue of *Survey Methodology*, Estevao and Särndal consider the problem of calibration estimation in the context of two-phase sampling. The contributions of the paper include the choice of initial weights in the calibration procedure as well as the important problem of variance estimation. New variance estimators are proposed and results from a simulation study show that the proposed variance estimators are more efficient than the traditional ones.

Next, Li and Valliant investigate the problem of the detection of influential units in linear regression analysis of survey data. They first give an expression for the hat matrix and its associated leverages (diagonal of the hat matrix) when a weighted least squares technique is used to estimate model parameters. They then propose a decomposition of the leverages and highlight that the leverage for a given unit can be large when either its survey weight is large or its vector of explanatory variables is far from the center. They illustrate the effect of influential units on both ordinary and weighted least squares using a numerical example.

Beaumont and Bocci propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The technique uses model-based test statistics that incorporate the survey weights and can usually be obtained easily using standard software packages. Using a simulation study the authors show that the proposed method performs similarly to the Rao-Scott procedure, and better than the Wald and Bonferroni procedures when testing hypotheses about a vector of linear regression model parameters.

The paper by Park, Choi and Choi present an interesting approach to nonresponse. Studies have shown that the voting behaviour of the undecided voters can have a significant impact on the final result of an election and that by considering these undecided voters, the accuracy of election forecasting can be improved. The authors present two Bayesian models whose priors depend on information from both respondents and undecided. They analyze an incomplete two-way contingency table using four sets of data from the 1998 Ohio state polls to illustrate how to use and interpret estimation results for the elections.

Ghosh, Kim, Sinha, Maiti, Katzoff and Parsons develop hierarchical and empirical Bayes methods for estimation of proportions in small domains using unit-level models. They propose a hierarchical Bayes analogue of the generalized linear mixed model to obtain posterior means and posterior standard errors of the population small domain proportions. Using an approach based on the theory of optimal estimating functions, they also obtain empirical Bayes estimators and corresponding asymptotic mean square error estimators. The methods are illustrated using data from the National Health Interview Survey (NHIS) to obtain small domain estimates of the proportions of Asians without health insurance.

In the McElroy and Holan paper, the problem of testing for residual seasonality in seasonally adjusted data is investigated. The authors propose a statistical significance test for peaks in the spectral density of the time series under consideration that is indicative of seasonality. The theory of the proposed method developed and is illustrated and compared with existing methods through both simulation and empirical studies.

Gabler and Lahiri provide a model-assisted justification of the traditional interviewer variance formula for equal probability sampling with no spatial clustering. They then obtain, in the context of a complex sampling design, a definition of interviewer variability that appropriately accounts for unequal probabilities of selection and spatial clustering. They also propose a decomposition of total effects into effects due to weighting, spatial clustering and interviewers. Their results can help to more effectively understand and control sources of variability.

In their paper, Schouten, Cobben and Bethlehem investigate the problem of assessing the similarity between the response to a survey and the sample or population under investigation. They propose a representativeness indicator to replace response rates as a quality indicator for the impact of nonresponse bias. This indicator, called the R-indicator, is shown to be somewhat related to Cramer's V measure for the association between response and auxiliary variables. In fact, the R-indicator is better viewed as a lack of association measure since a weaker association implies that there is no evidence that nonresponse has affected the composition of the observed data. The theoretical properties of the proposed indicator are developed and it is illustrated through empirical studies.

Finally, in his article, Chauvet addresses the issue of balanced sampling when sizes in each stratum are too small for exact balancing. The author proposes an algorithm adapted to the Cube method, which guarantees balancing at the population level. A simulation study confirmed that the proposed method performed well.

Harold Mantel, Deputy Editor

A new face on two-phase sampling with calibration estimators

Victor M. Estevao and Carl-Erik Särndal¹

Abstract

This paper provides a framework for estimation by calibration in two-phase sampling designs. This work grew out of the continuing development of generalized estimation software at Statistics Canada. An important objective in this development is to provide a wide range of options for effective use of auxiliary information in different sampling designs. This objective is reflected in the general methodology for two-phase designs presented in this paper.

We consider the traditional two-phase sampling design. A phase-one sample is drawn from the finite population and then a phase-two sample is drawn as a sub-sample of the first. The study variable, whose unknown population total is to be estimated, is observed only for the units in the phase-two sample. Arbitrary sampling designs are allowed in each phase of sampling. Different types of auxiliary information are identified for the computation of the calibration weights at each phase. The auxiliary variables and the study variables can be continuous or categorical.

The paper contributes to four important areas in the general context of calibration for two-phase designs:

- (1) Three broad types of auxiliary information for two-phase designs are identified and used in the estimation. The information is incorporated into the weights in two steps: a phase-one calibration and a phase-two calibration. We discuss the composition of the appropriate auxiliary vectors for each step, and use a linearization method to arrive at the residuals that determine the asymptotic variance of the calibration estimator.
- (2) We examine the effect of alternative choices of starting weights for the calibration. The two "natural" choices for the starting weights generally produce slightly different estimators. However, under certain conditions, these two estimators have the same asymptotic variance.
- (3) We re-examine variance estimation for the two-phase calibration estimator. A new procedure is proposed that can improve significantly on the usual technique of conditioning on the phase-one sample. A simulation in section 10 serves to validate the advantage of this new method.
- (4) We compare the calibration approach with the traditional model-assisted regression technique which uses a linear regression fit at two levels. We show that the model-assisted estimator has properties similar to a two-phase calibration estimator.

Key Words: Auxiliary information; Two-phase regression estimator; Starting weights; Separate residual variance estimator; Combined residual variance estimator.

1. Introduction

The term *double sampling* refers to sampling designs whose common feature is a selection of two probability samples, denoted s_1 and s , both of them subsets of the finite population of interest, given by $U = \{1, \dots, k, \dots, N\}$. The sample s_1 is realized and observed prior to s . A typical study variable is denoted by y ; its value y_k is obtained only for the units $k \in s$. The objective is to estimate the population y -total $Y = \sum_U y_k$ (if A is a set of units, $A \subseteq U$, then we write \sum_A as a short form for $\sum_{k \in A}$ when there is no ambiguity).

Hidirolou (2001) discusses two types of double sampling, *nested* and *non-nested*. This paper focuses on the nested type, usually referred to as two-phase sampling: The phase-two sample s is a sub-sample from the phase-one sample s_1 drawn from U , so $s \subseteq s_1 \subseteq U$.

Estimation for two-phase sampling has been examined in several earlier papers in a context where two kinds of auxiliary information are recognized and addressed by their

levels: At the population level, the total $\sum_U \mathbf{x}_{1k}$ is known, where \mathbf{x}_{1k} is a vector known for every $k \in s_1$; therefore, it is also known for every $k \in s$. At the level of the first sample, the vector value \mathbf{x}_{2k} is observed for every $k \in s_1$, and is thereby known for every $k \in s$; the total $\sum_U \mathbf{x}_{2k}$ is unknown but can be estimated without design bias at the s_1 -level. Two arguments are found in the literature for incorporating these two types of auxiliary information in estimating $Y = \sum_U y_k$: the regression fit argument and the calibration argument. Under certain conditions they can lead to identical estimators, but this is not so in general.

The regression fit argument prevails in Särndal and Swensson (1987), Särndal, Swensson and Wretman (1992), Sitter (1997), Hidirolou and Särndal (1998), Axelson (1998) and Hidirolou, Rao and Haziza (2006). The calibration approach in Deville and Särndal (1992) was applied to two-phase sampling by Dupont (1995). She compares the resulting calibration estimators with those obtained from the regression approach. For the same auxiliary information, the two approaches may not give

1. Victor M. Estevao, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. E-mail: victor.estevao@statcan.gc.ca; Carl-Erik Särndal, professor. E-mail: carl.sarndal@rogers.com.

identical estimators, although in practice the difference is likely to be of little consequence. Resampling for two-phase variance estimation is considered in Kott and Stukel (1997). Estevao and Särndal (2002) focus on the calibration argument and distinguish ten different ways to use all or part of the information available at the two levels. The present paper also focuses on the calibration approach. It extends earlier work by recognizing three (rather than two) types of auxiliary information, each having different characteristics.

In the regression approach, it is natural to fit two linear least squares regressions. One set of regression-predicted y -values are produced for $k \in s_1$ using both \mathbf{x}_{1k} and \mathbf{x}_{2k} as predictors; another set is produced for $k \in s_1$ using only the vector \mathbf{x}_{1k} as predictor. Both sets of predicted y -values, as well as the known total $\sum_U \mathbf{x}_{1k}$, are used to build the regression-type estimator of Y , in the manner described in section 9.

The calibration approach is motivated by two factors: To create a set of weights that are consistent with known or estimated totals for the auxiliary variables and to reduce the variance of the estimates made for the study variable(s). We want the weights w_k in $\hat{Y}_{2p} = \sum_s w_k y_k$ to achieve consistency with the total $\sum_U \mathbf{x}_{1k}$ known at the level of the population and/or with an (approximately) unbiased estimate, made at the level of the phase-one sample, of the unknown $\sum_U \mathbf{x}_{2k}$. Since y is observed only at the ultimate level (the phase-two sample), consistency “at higher levels” on important auxiliary variables will often significantly reduce the variance of $\hat{Y}_{2p} = \sum_s w_k y_k$. We can distinguish two steps in the process leading to the weights w_k , a phase-one calibration and a phase-two calibration.

The two-phase sampling design is as follows: From the finite population of units $U = \{1, 2, \dots, k, \dots, N\}$ we select a phase-one sample s_1 . The known positive inclusion probability of unit k is $\pi_{1k} = \Pr(k \in s_1)$, and the phase-one design weight is $a_{1k} = 1/\pi_{1k}$. Certain variables may be observed for the units $k \in s_1$. Then, conditionally on s_1 , we select a phase-two sample s from s_1 . The known and positive conditional inclusion probability of k is $\pi_{2k} = \Pr(k \in s | s_1)$ for $k \in s_1$, and the conditional phase-two design weight is $a_{2k} = 1/\pi_{2k}$. (to keep the notation simple, we use π_{2k} and a_{2k} rather than the more suggestive $\pi_{2k|s_1}$ and $a_{2k|s_1}$; it should be kept in mind that both π_{2k} and a_{2k} are conditional on the phase-one sample s_1). The combined or double-expansion design weight is $a_k = a_{1k}a_{2k}$ for $k \in s$. The analysis of the estimators in this article is design based. The term “(approximately) unbiased” means “(approximately) design unbiased.” We assume mild conditions on the population and the two sampling designs, permitting us to discard lower order terms in the analysis of our estimators when the expected sizes of the phase-one and phase-two samples are sufficiently large.

The double-expansion estimator $\sum_s a_k y_k$ is unbiased for $Y = \sum_U y_k$. We can produce more efficient estimators by taking into account the available auxiliary information. Three types or sets of auxiliary variables (called x -variables) can be distinguished for two-phase sampling designs. These are denoted by \mathcal{X}^\oplus , \mathcal{X}^\dagger and \mathcal{X}° . Their information characteristics are specified in the following table.

Table 1.1
Sets of auxiliary variables for calibration in two-phase sampling

Set of auxiliary variables	Auxiliary variable total over U	Unit variable values for $k \in s_1$	Unit variable values for $k \in s$
\mathcal{X}^\oplus	known	known	known
\mathcal{X}^\dagger	known	unknown	known
\mathcal{X}°	unknown	known	known

Each set may contain any number of x -variables. The three sets are mutually exclusive. The properties in the last three columns apply to every x -variable in the corresponding set. All x -variables used for calibration belong to one of these three sets.

2. Phase-one calibration

For the phase-one calibration, we use a vector \mathbf{x}_{1k} of auxiliary variables selected from the set \mathcal{X}^\oplus . While it is natural to let \mathbf{x}_{1k} consist of all the variables in \mathcal{X}^\oplus , the general presentation here allows us to define \mathbf{x}_{1k} to include some or even none of the variables in \mathcal{X}^\oplus . The phase-one calibration weights w_{1k} are derived by modifying the phase-one starting weights a_{1k} subject to the calibration constraint $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. In our formulation, the calibration weights are given for $k \in s_1$ as

$$w_{1k} = a_{1k} \left\{ 1 + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}' \right)^{-1} \mathbf{z}_{1k} \right\} \quad (2.1)$$

where $\mathbf{X}_1 = \sum_U \mathbf{x}_{1k}$, $\hat{\mathbf{X}}_1 = \sum_{s_1} a_{1k} \mathbf{x}_{1k}$ and \mathbf{z}_{1k} is an instrumental vector of the same dimension as \mathbf{x}_{1k} . It replaces $\mathbf{x}_{1k}/\sigma_{1k}^2$ in the form of the model-assisted estimator described by Särndal, Swensson, Wretman (1992), and permits a more general specification of the calibration weights. The use of an instrumental vector is discussed in Estevao and Särndal (2000) and Deville (2002). Here and in the following, we always assume the invertibility of matrices such as the one over s_1 in (2.1) and those (over s and U) appearing later.

3. Phase-two calibration

We use a vector \mathbf{x}_k of auxiliary variables to produce a set of phase-two (or final) calibration weights w_k . They are

used to calculate $\hat{Y}_{2p} = \sum_s w_k y_k$ as our estimator of $Y = \sum_U y_k$. The vector $\mathbf{x}_k = (\mathbf{x}'_{k(l)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$ has three components, as described below. No auxiliary variable can appear in more than one of the three vector components. These three components have different roles in the setup of the phase-two calibration equation $\sum_s w_k \mathbf{x}_k = \mathbf{X}$ and in the determination of the phase-two calibration weights.

The variables in the vector $\mathbf{x}_{k(l)}$ are selected from among those in the set $\mathcal{X}^{\oplus} \cup \mathcal{X}^{\dagger}$. This means that the total $\sum_U \mathbf{x}_{k(l)}$ is known and can be included in \mathbf{X} . Variables in \mathbf{x}_{lk} are allowed to reoccur in $\mathbf{x}_{k(l)}$, and this is usually preferable in order to reduce the variance of the estimator. We can specify $\mathbf{x}_{k(l)} = \mathbf{x}_{lk}$, but our framework permits $\mathbf{x}_{k(l)}$ to include variables from \mathcal{X}^{\dagger} . This allows us to use variables with known population totals in situations where the variables are too expensive to collect for a large phase-one sample s_1 but are observable for the smaller phase-two sample s . These variables are excluded from the phase-one calibration because they are unavailable for $k \in s_1$.

The variables in $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{k(a)}$ are selected from among those in the set $\mathcal{X}^{\oplus} \cup \mathcal{X}^{\dagger} \cup \mathcal{X}^{\circ}$ provided they are not already included in $\mathbf{x}_{k(l)}$. The variables in $\mathbf{x}_{k(w)}$ are those for which we want to satisfy the phase-two calibration equation $\sum_s w_k \mathbf{x}_{k(w)} = \sum_{s_1} w_{lk} \mathbf{x}_{k(w)}$, where the right-hand side is approximately unbiased for $\sum_U \mathbf{x}_{k(w)}$. The variables in $\mathbf{x}_{k(a)}$ are those for which we want to satisfy the phase-two calibration equation $\sum_s w_k \mathbf{x}_{k(a)} = \sum_{s_1} a_{lk} \mathbf{x}_{k(a)}$. Here, the right-hand side is unbiased for $\sum_U \mathbf{x}_{k(a)}$. The inclusion of both $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{k(a)}$ in the definition of \mathbf{x}_k allows us to calibrate on one or both of these vectors and provides a general framework for producing different estimators from the phase-two calibration.

The phase-two calibration equation is $\sum_s w_k \mathbf{x}_k = \mathbf{X}$, where \mathbf{X} is the stacked auxiliary vector

$$\mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_{k(l)} \\ \sum_{s_1} w_{lk} \mathbf{x}_{k(w)} \\ \sum_{s_1} a_{lk} \mathbf{x}_{k(a)} \end{pmatrix}. \quad (3.1)$$

A specific variable can only occur once in \mathbf{x}_k . Otherwise, the calibration equation may be inconsistent and admit no solution.

The starting weights for the phase-two calibration are denoted by a_k^* for $k \in s$. There is more than one reasonable choice for the a_k^* . We consider two alternatives, both of which seem natural: (1) $a_k^* = a_k = a_{lk} a_{2k}$, and (2) $a_k^* = w_{lk} a_{2k}$, where w_{lk} is the phase-one calibration weight given by (2.1).

Given the starting weights a_k^* , we determine final weights w_k subject to the calibration equation $\sum_s w_k \mathbf{x}_k = \mathbf{X}$. These final weights are given for $k \in s$ by

$$w_k = a_k^* \left\{ 1 + (\mathbf{X} - \hat{\mathbf{X}})' \left(\sum_s a_k^* \mathbf{z}_k \mathbf{x}_k' \right)^{-1} \mathbf{z}_k \right\} \quad (3.2)$$

where $\hat{\mathbf{X}} = \sum_s a_k^* \mathbf{x}_k$ is an unbiased or approximately unbiased estimator of \mathbf{X} , depending on the composition of \mathbf{x}_k . The instrumental variable \mathbf{z}_k has the same dimension as \mathbf{x}_k . The vectors \mathbf{z}_{lk} and \mathbf{z}_k are assumed to be fixed functions of \mathbf{x}_{lk} and \mathbf{x}_k . How to choose \mathbf{z}_{lk} and \mathbf{z}_k is a topic we leave for others to address.

4. Comparison of two options for the starting weights

The objective in this section is to analyze how the final weights w_k in $\hat{Y}_{2p} = \sum_s w_k y_k$ depend on the specification of the starting weights a_k^* in (3.2). We consider two distinct cases based on whether or not the auxiliary variables \mathbf{x}_k are used for the phase-two calibration. When we carry out the phase-two calibration, the two different choices for starting weights generally lead to different estimators. We show that these estimators are asymptotically equivalent under certain conditions, commonly found in practice. When we have no phase-two calibration, the two choices for starting weights lead to two other estimators that are usually less efficient than those obtained by performing the phase-two calibration.

4.1 Estimators with phase-two calibration ($\mathbf{x}_k \neq \phi$)

As noted previously, there are two alternatives for the starting weights a_k^* in (3.2): (1) $a_k^* = a_k = a_{lk} a_{2k}$, and (2) $a_k^* = w_{lk} a_{2k}$, where w_{lk} is the phase-one calibration weight given by (2.1). We now provide a detailed analysis of the form of the estimator under these two choices. In this subsection, we look at the more interesting case where we perform the phase-two calibration ($\mathbf{x}_k \neq \phi$). In the next subsection, we consider what happens when we do not carry out the phase-two calibration ($\mathbf{x}_k = \phi$).

Our procedure is as follows. First, we derive the linearized (asymptotic) form of \hat{Y}_{2p} based on the general starting weights a_k^* . Then we substitute the two choices for a_k^* in this expression. We determine \hat{Y}_{2p} based on the starting weights $a_k^* = a_k = a_{lk} a_{2k}$. We denote this estimator by $\hat{Y}_{2p,a}$ and derive its linearized form, $\hat{Y}_{2p,a \text{ lin}}$. Similarly, we obtain \hat{Y}_{2p} based on the starting weights $a_k^* = w_{lk} a_{2k}$. We refer to this estimator as $\hat{Y}_{2p,w}$ and derive its linearized form, $\hat{Y}_{2p,w \text{ lin}}$. These two forms are slightly different but we prove in Result 4.2 that $\hat{Y}_{2p,a \text{ lin}} = \hat{Y}_{2p,w \text{ lin}}$ under certain conditions.

We start by inserting the weights w_k into $\hat{Y}_{2P} = \sum_S w_k y_k$ and writing the estimator as

$$\begin{aligned} \hat{Y}_{2P} &= \sum_U \mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x})(t)} + \sum_{s_1} w_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)} \\ &+ \sum_{s_1} a_{1k} \mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + \sum_S a_k^* e_{(y; \mathbf{x})k} \\ &+ (\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y; \mathbf{x})}^* - \mathbf{B}_{(y; \mathbf{x})}) \end{aligned} \quad (4.1)$$

where $\hat{\mathbf{B}}_{(y; \mathbf{x})}^* = (\sum_S a_k^* \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_S a_k^* \mathbf{z}_k y_k$, $\mathbf{B}_{(y; \mathbf{x})} = (\sum_U \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_U \mathbf{z}_k y_k$ and $\mathbf{B}_{(y; \mathbf{x})} = (\mathbf{B}'_{(y; \mathbf{x})(t)}, \mathbf{B}'_{(y; \mathbf{x})(w)}, \mathbf{B}'_{(y; \mathbf{x})(a)})'$ is the partitioning corresponding to $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$. Our subscript notation of the form $(\mathbf{v}_1; \mathbf{v}_2)$ identifies the variables in the regression. The term \mathbf{v}_2 refers to the independent variables and \mathbf{v}_1 identifies the dependent variable or variables. For simplicity, the instrumental vectors \mathbf{z}_{1k} and \mathbf{z}_k are not included in the notation.

The term $e_{(y; \mathbf{x})k} = y_k - \mathbf{x}'_k \mathbf{B}_{(y; \mathbf{x})}$ is defined for $k \in U$. Note that although $e_{(y; \mathbf{x})k}$ looks like a regression residual, it does not arise as the result of fitting a proper regression model. We then develop the term $\sum_{s_1} w_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)}$ in (4.1) by inserting expression (2.1) for w_{1k} and making use of the phase-one calibration equation $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. We obtain

$$\begin{aligned} \sum_{s_1} w_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)} &= \mathbf{X}'_1 \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} + \sum_{s_1} a_{1k} e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k} \\ &+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) \end{aligned} \quad (4.2)$$

where $\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} = (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} y_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)}$ converges in probability to (and is approximately unbiased for) $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_U \mathbf{z}_{1k} y_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)}$, and $e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k} = \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)} - \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}$ is defined for $k \in U$.

We can interpret $e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k}$ as a residual arising from a population fit based on a generalized regression of $\mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x})(w)}$ as the dependent variable and \mathbf{x}_{1k} as the predictor vector. Replacing expression (4.2) into expression (4.1) for \hat{Y}_{2P} leads to

$$\begin{aligned} \hat{Y}_{2P} &= \sum_U (\mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x})(t)} + \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) \\ &+ \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k}) \\ &+ \sum_S a_k^* e_{(y; \mathbf{x})k} \\ &+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) \\ &+ (\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y; \mathbf{x})}^* - \mathbf{B}_{(y; \mathbf{x})}). \end{aligned} \quad (4.3)$$

The following result establishes the relationship between the estimators obtained for the two choices of starting weights.

Result 4.1: The linearized forms of \hat{Y}_{2Pa} and \hat{Y}_{2Pw} are related by the equation $\hat{Y}_{2Pw} = \hat{Y}_{2Pa} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)})$.

Proof

We consider expression (4.3) under the two possible choices for a_k^* . First, with $a_k^* = a_k = a_{1k} a_{2k}$ we obtain \hat{Y}_{2Pa} given by

$$\begin{aligned} \hat{Y}_{2Pa} &= \sum_U (\mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x})(t)} + \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) \\ &+ \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k}) \\ &+ \sum_S a_k e_{(y; \mathbf{x})k} \\ &+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) \\ &+ (\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y; \mathbf{x})} - \mathbf{B}_{(y; \mathbf{x})}). \end{aligned} \quad (4.4)$$

The term $\hat{\mathbf{B}}_{(y; \mathbf{x})} = (\sum_S a_k \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_S a_k \mathbf{z}_k y_k$ converges in probability to (and is approximately unbiased for) $\mathbf{B}_{(y; \mathbf{x})} = (\sum_U \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_U \mathbf{z}_k y_k$. The first term is constant and does not contribute to the variance of \hat{Y}_{2Pa} . The next two terms are random quantities, defined as sums over s_1 and S respectively. The last two terms are products of differences with zero or almost zero expectation. As for the product $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)})$, both differences are functions of the phase-one sample s_1 . We know that $\hat{\mathbf{X}}_1$ is unbiased for \mathbf{X}_1 and $\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}$ is approximately unbiased for $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}$. Under fairly general conditions, $N^{-1}(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) = O_p(n_1^{-1})$, where n_1 is the expected size of s_1 , assumed sufficiently large. By a similar reasoning, $N^{-1}(\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y; \mathbf{x})} - \mathbf{B}_{(y; \mathbf{x})}) = O_p(n^{-1})$, where n is the expected size of S , also assumed sufficiently large. Consequently, we can drop the last two terms of (4.4), because they are of lower order than the preceding terms: $N^{-1} \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k})$ is $O_p(n_1^{-1/2})$ and $N^{-1} \sum_S a_k e_{(y; \mathbf{x})k}$ is $O_p(n^{-1/2})$. The first three terms define the linearized form of \hat{Y}_{2Pa} ,

$$\begin{aligned} \hat{Y}_{2Pa \text{ lin}} &= \sum_U (\mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x})(t)} + \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)}) \\ &+ \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)}; \mathbf{x}_1)k}) \\ &+ \sum_S a_k e_{(y; \mathbf{x})k}. \end{aligned} \quad (4.5)$$

Now let us consider expression (4.3) under the second choice, $a_k^* = w_{1k} a_{2k}$. This leads to \hat{Y}_{2Pw} given by

$$\begin{aligned}
\hat{Y}_{2Pw} = & \sum_U (\mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x})(t)} + \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)}) \\
& + \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + e_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)k}) \\
& + \sum_s a_k e_{(y; \mathbf{x})k} \\
& + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k} \right)^{-1} \sum_s a_k \mathbf{z}_{1k} e_{(y; \mathbf{x})k} \\
& + (\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y; \mathbf{x})}^w - \mathbf{B}_{(y; \mathbf{x})}) \quad (4.6)
\end{aligned}$$

where $\hat{\mathbf{B}}_{(y; \mathbf{x})}^w = (\sum_s w_{1k} a_{2k} \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_s w_{1k} a_{2k} \mathbf{z}_k y_k$ and $\hat{\mathbf{X}} = \sum_s w_{1k} a_{2k} \mathbf{x}_k$. The first three terms of \hat{Y}_{2Pw} are the same as those found in expression (4.4) for \hat{Y}_{2Pa} . The fourth and fifth terms differ from their counterparts in (4.4). Although $\hat{\mathbf{B}}_{(y; \mathbf{x})}^w$ and $\hat{\mathbf{X}}$ are functions of the phase-one calibration weights w_{1k} , we do not need to replace them in $\hat{\mathbf{B}}_{(y; \mathbf{x})}^w$ and $\hat{\mathbf{X}}$ in the fifth term; this would simply split the lower order term $(\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y; \mathbf{x})}^w - \mathbf{B}_{(y; \mathbf{x})})$ into other lower order terms. Therefore, we can drop the fifth term of (4.6) when the sample sizes are sufficiently large. The fourth term can be written as follows.

$$\begin{aligned}
(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k} \right)^{-1} \sum_s a_k \mathbf{z}_{1k} e_{(y; \mathbf{x})k} \\
= (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})}) \\
+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1} - \mathbf{B}_{(y; \mathbf{x}_1)}) \\
- (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)}) \mathbf{B}_{(y; \mathbf{x})} \\
+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k} \right)^{-1} \\
\left(\sum_s a_k \mathbf{z}_{1k} e_{(y; \mathbf{x})k} - \sum_{s_1} a_{1k} \mathbf{z}_{1k} e_{(y; \mathbf{x})k} \right). \quad (4.7)
\end{aligned}$$

The quantities in this expression are defined as follows: $\hat{\mathbf{B}}_{(\mathbf{x}; \mathbf{x}_1)} = (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_k$ and $\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1} = (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} y_k$. The statistic $\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1}$ can not be computed from the phase-one sample because the values y_k are only known for $k \in s$. It is implicitly defined for the purpose of determining the linearized form. We can define such a construct in the same manner as $\hat{\mathbf{B}}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)}$ is a function of the unknown quantity $\mathbf{B}_{(y; \mathbf{x})(w)}$. Now $\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1}$ is approximately unbiased for its corresponding population quantity $\mathbf{B}_{(y; \mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_U \mathbf{z}_{1k} y_k$. Similarly, $\hat{\mathbf{B}}_{(\mathbf{x}; \mathbf{x}_1)}$ is approximately unbiased for $\mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_U \mathbf{z}_{1k} \mathbf{x}'_k$. As before, we can argue that the last three terms of (4.7) are of lower order than the first term $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})})$, which provides the linear approximation. The substitution of this term into (4.6) leads to the linearized form of \hat{Y}_{2Pw} ,

$$\begin{aligned}
\hat{Y}_{2Pw \text{ lin}} = & \sum_U (\mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x})(t)} + \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)}) \\
& + \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x})(a)} + e_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)k}) \\
& + \sum_s a_k e_{(y; \mathbf{x})k} \\
& + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})}). \quad (4.8)
\end{aligned}$$

Comparing (4.5) with (4.8), we see that $\hat{Y}_{2Pw \text{ lin}} = \hat{Y}_{2Pa \text{ lin}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})})$ as stated in the result. This completes the proof of result 4.1.

Result 4.1 shows that in general, the linearized forms of \hat{Y}_{2Pw} and \hat{Y}_{2Pa} are not the same. However, they are the same under certain conditions. Let us consider the case of nested calibration (not to be confused with nested sampling), meaning that \mathbf{x}_k includes \mathbf{x}_{1k} . Then \mathbf{x}_k is of the form $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ where the vector \mathbf{x}_{+k} is composed of the remaining variables. We now state and prove the following result.

Result 4.2: If $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ and $\mathbf{z}_k = (\mathbf{z}'_{1k}, \mathbf{z}'_{+k})'$ then $\hat{Y}_{2Pw \text{ lin}} = \hat{Y}_{2Pa \text{ lin}}$ and \hat{Y}_{2Pw} and \hat{Y}_{2Pa} are asymptotically equivalent.

Proof

The proof follows from result 4.1 by showing $\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})} = \mathbf{0}$ under the specified conditions. We have

$$\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})} = \left(\sum_U \mathbf{z}_{1k} \mathbf{x}'_{1k} \right)^{-1} \left(\sum_U \mathbf{z}_{1k} h_k \right)$$

where $h_k = y_k - \mathbf{x}'_{1k} (\sum_U \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} (\sum_U \mathbf{z}_{1k} y_k)$. Since $\sum_U \mathbf{z}_{1k} h_k = \mathbf{0}$ and we assume $\mathbf{z}_k = (\mathbf{z}'_{1k}, \mathbf{z}'_{+k})'$, it follows $\sum_U \mathbf{z}_{1k} h_k = \mathbf{0}$ and $\mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})} = \mathbf{0}$. Therefore from result 4.1, $\hat{Y}_{2Pw \text{ lin}} = \hat{Y}_{2Pa \text{ lin}}$. Since their linear forms are the same, \hat{Y}_{2Pw} and \hat{Y}_{2Pa} are asymptotically equivalent estimators.

Interestingly, Result 4.2 only requires that we include \mathbf{x}_{1k} somewhere within \mathbf{x}_k . Obviously, it makes sense to include \mathbf{x}_{1k} within the component $\mathbf{x}_{k(t)}$ of \mathbf{x}_k because the \mathbf{x}_{1k} -totals are known. However, we obtain the same asymptotic result as long as all variables in \mathbf{x}_{1k} are included somewhere in $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$. In practice, we often find $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$ with $\mathbf{z}_{1k} = \mathbf{x}_{1k}$ and $\mathbf{z}_k = \mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ where \mathbf{x}_{+k} is the vector for the remaining variables $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{k(a)}$. This satisfies the requirements for the asymptotic equivalence of \hat{Y}_{2Pa} and \hat{Y}_{2Pw} .

To study the properties of \hat{Y}_{2Pa} and \hat{Y}_{2Pw} we work with their linearized forms given respectively by (4.5) and (4.8). With appropriate definitions for the residuals e_{0k} , e_{1k} and e_{2k} , we can represent $\hat{Y}_{2Pa \text{ lin}}$ and $\hat{Y}_{2Pw \text{ lin}}$ as the sum of three terms: a constant term $\sum_U e_{0k}$, a phase-one expansion term $\sum_{s_1} a_{1k} e_{1k}$, and a double-expansion term $\sum_s a_k e_{2k}$,

$$\hat{Y}_{2P \text{ lin}} = \sum_U e_{0k} + \sum_{s_1} a_{1k} e_{1k} + \sum_s a_k e_{2k}. \quad (4.9)$$

This makes (4.9) a suitable starting point for studying the bias and the asymptotic variance of the two estimators $\hat{Y}_{2P a}$ and $\hat{Y}_{2P w}$.

For the linearized form $\hat{Y}_{2P a \text{ lin}}$ given by (4.5), the three residual quantities are defined as follows for $k \in U$:

$$\begin{aligned} e_{0k} &= \mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x}(t))} + \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)} \\ e_{1k} &= \mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x}(a))} + \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x}(w))} \\ &\quad - \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)} \\ e_{2k} &= y_k - \mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x}(t))} - \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x}(w))} \\ &\quad - \mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x}(a))}. \end{aligned} \quad (4.10)$$

Note that e_{2k} is simply $e_{(y; \mathbf{x})k}$. Similarly, for $\hat{Y}_{2P w \text{ lin}}$ given by (4.8), the residuals have the following definitions for $k \in U$:

$$\begin{aligned} e_{0k} &= \mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x}(t))} \\ &\quad + \mathbf{x}'_{1k} (\mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)} + \mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})}) \\ e_{1k} &= \mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x}(a))} + \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x}(w))} \\ &\quad - \mathbf{x}'_{1k} (\mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)} + \mathbf{B}_{(y; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}; \mathbf{x}_1)} \mathbf{B}_{(y; \mathbf{x})}) \\ e_{2k} &= y_k - \mathbf{x}'_{k(t)} \mathbf{B}_{(y; \mathbf{x}(t))} \\ &\quad - \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x}(w))} - \mathbf{x}'_{k(a)} \mathbf{B}_{(y; \mathbf{x}(a))}. \end{aligned} \quad (4.11)$$

Note that in both cases, $e_{0k} + e_{1k} + e_{2k} = y_k$ for every k , and hence $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y$. This additivity allows us to prove in section 5 that $\hat{Y}_{2P a}$ and $\hat{Y}_{2P w}$ are approximately unbiased. To save space, we concentrate on the properties of $\hat{Y}_{2P a}$ in the remaining sections. However, the analysis is similar for $\hat{Y}_{2P w}$ and the method for variance estimation proposed in section 7 can also be used for this estimator.

4.2 Estimators without the phase-two calibration ($\mathbf{x}_k = \phi$)

If there is no phase-two calibration ($\mathbf{x}_k = \phi$), then $w_k = a_k^*$. Accordingly, the final weights are either $w_k = a_k = a_{1k} a_{2k}$ or $w_k = w_{1k} a_{2k}$. The first alternative gives the double-expansion estimator $\sum_s a_k y_k$. The second produces a different estimator that is usually more efficient. However, both of these are generally inefficient compared to the estimators obtained by carrying out the phase-two

calibration. The linearized form of the two-phase estimator with $w_k = w_{1k} a_{2k}$ is obtained by writing it as follows.

$$\begin{aligned} \hat{Y}_{2P} &= \mathbf{X}'_1 \mathbf{B}_{(y; \mathbf{x}_1)} - \hat{\mathbf{X}}'_1 \mathbf{B}_{(y; \mathbf{x}_1)} + \sum_s a_k y_k \\ &\quad + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1} - \mathbf{B}_{(y; \mathbf{x}_1)}) \\ &\quad + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k} \right)^{-1} \\ &\quad \left(\sum_s a_k \mathbf{z}_{1k} y_k - \sum_{s_1} a_{1k} \mathbf{z}_{1k} y_k \right). \end{aligned} \quad (4.12)$$

The terms $\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1}$ and $\mathbf{B}_{(y; \mathbf{x}_1)}$ were defined in the previous section. When the samples are sufficiently large, $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} (\sum_s a_k \mathbf{z}_{1k} y_k - \sum_{s_1} a_{1k} \mathbf{z}_{1k} y_k)$ and $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s_1} - \mathbf{B}_{(y; \mathbf{x}_1)})$ are of lower order and can be ignored. This leads to the linearized form of this estimator.

$$\hat{Y}_{2P \text{ lin}} = \mathbf{X}'_1 \mathbf{B}_{(y; \mathbf{x}_1)} - \hat{\mathbf{X}}'_1 \mathbf{B}_{(y; \mathbf{x}_1)} + \sum_s a_k y_k. \quad (4.13)$$

We can also write this linearized form as a sum (4.9) of three residual terms, with the residuals e_{0k} , e_{1k} and e_{2k} having following definitions for $k \in U$.

$$\begin{aligned} e_{0k} &= \mathbf{x}'_{1k} \mathbf{B}_{(y; \mathbf{x}_1)} \\ e_{1k} &= -\mathbf{x}'_{1k} \mathbf{B}_{(y; \mathbf{x}_1)} \\ e_{2k} &= y_k. \end{aligned} \quad (4.14)$$

These residuals show a resemblance to those given by (4.10) if we set $\mathbf{x}_k = \phi$ and remove $\mathbf{B}_{(y; \mathbf{x})}$. Note how $\mathbf{B}_{(y; \mathbf{x}_1)}$ has the same role as $\mathbf{B}_{(\mathbf{x} \mathbf{B}_{(w)}; \mathbf{x}_1)}$ in (4.10). As before, $e_{0k} + e_{1k} + e_{2k} = y_k$ for every k , and hence $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y$.

The double-expansion estimator is a special case of this estimator when we also have $\mathbf{x}_{1k} = \phi$. This means that $\mathbf{B}_{(y; \mathbf{x}_1)}$ is not defined. The corresponding definitions for e_{0k} , e_{1k} and e_{2k} are simply $e_{0k} = 0$, $e_{1k} = 0$ and $e_{2k} = y_k$ for $k \in U$.

In the following sections, we examine the bias and variance of the two-phase calibration estimator $\hat{Y}_{2P a}$ and we propose a new method for estimation of variance. We can derive corresponding results when there is no phase-two calibration because the residuals for these two groups of estimators have similar properties and linearized form. The only difference occurs in the estimation of variance. We use the same variance estimator (as described in section 7) but the residuals are estimated by using $\hat{e}_{1k} = -\mathbf{x}'_{1k} \hat{\mathbf{B}}_{(y; \mathbf{x}_1) s}$ where $\hat{\mathbf{B}}_{(y; \mathbf{x}_1) s} = (\sum_s a_k \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_s a_k \mathbf{z}_{1k} y_k$, and $\hat{e}_{2k} = y_k$.

5. Bias and variance of the two-phase calibration estimator \hat{Y}_{2Pa}

The two-phase calibration estimator $\hat{Y}_{2Pa} = \sum_s w_k y_k$ is approximately unbiased for $Y = \sum_U y_k$. To show this, we derive the expectation of the linearized form given by (4.9) via the usual method of conditioning on the phase-one sample s_1 . We have $E(\sum_s a_k e_{2k}) = E_{s_1} E_{s_1|s_1}(\sum_s a_k e_{2k}) = E_{s_1}(\sum_{s_1} a_{1k} e_{2k}) = \sum_U e_{2k}$, $E_{s_1}(\sum_{s_1} a_{1k} e_{1k}) = \sum_U e_{1k}$, and $\sum_U e_{0k}$ is a constant term, so

$$E(\hat{Y}_{2Pa \text{ lin}}) = \sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y.$$

This shows that $\hat{Y}_{2Pa \text{ lin}}$ is unbiased for Y . By (4.4), $\hat{Y}_{2Pa} = \hat{Y}_{2Pa \text{ lin}} + R$, so the bias of \hat{Y}_{2Pa} equals the expectation of R , which is the sum of the two lower order terms $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)})$ and $(\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}_{(y; \mathbf{x})} - \mathbf{B}_{(y; \mathbf{x})})$. As pointed out in section 4, each of these terms has expectation close to zero. It follows that \hat{Y}_{2Pa} is approximately unbiased for Y .

The variance of $\hat{Y}_{2Pa} = \sum_s w_k y_k$ is closely approximated by the variance of the linearized form $\hat{Y}_{2Pa \text{ lin}}$ given by (4.9) with residuals defined by (4.10). Its first term, $\sum_U e_{0k}$, is constant and does not contribute to the variance. Therefore,

$$V(\hat{Y}_{2Pa \text{ lin}}) = V\left(\sum_{s_1} a_{1k} e_{1k} + \sum_s a_k e_{2k}\right). \quad (5.1)$$

We use (5.1) as the starting point for deriving a variance estimator for $\hat{Y}_{2Pa \text{ lin}}$. Two different approaches can be used and it is of interest to compare them. The one in section 7 is new and more interesting because it produces a more efficient variance estimator than the one in section 8, derived by the traditional technique of conditioning on the phase-one sample s_1 . The residuals e_{1k} and e_{2k} given by (4.10) play an important role in both derivations.

6. Preliminaries for variance estimation

Our objective is to estimate the variance $V(\hat{Y}_{2Pa \text{ lin}})$ given by (5.1). This is done in sections 7 and 8 by two different arguments. The residuals e_{1k} and e_{2k} are defined for all $k \in U$ but they can not be computed. They must be replaced by estimates \hat{e}_{1k} and \hat{e}_{2k} . These estimates, formed in the image of (4.10) are

$$\begin{aligned} \hat{e}_{1k} &= \mathbf{x}'_{k(a)} \hat{\mathbf{B}}_{(y; \mathbf{x}(a))} + \mathbf{x}'_{k(w)} \hat{\mathbf{B}}_{(y; \mathbf{x}(w))} \\ &\quad - \mathbf{x}'_{1k} \hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)} \quad \text{for } k \in s_1 \\ \hat{e}_{2k} &= y_k - \mathbf{x}'_k \hat{\mathbf{B}}_{(y; \mathbf{x})} \\ &= y_k - \mathbf{x}'_{k(t)} \hat{\mathbf{B}}_{(y; \mathbf{x}(t))} - \mathbf{x}'_{k(w)} \hat{\mathbf{B}}_{(y; \mathbf{x}(w))} \\ &\quad - \mathbf{x}'_{k(a)} \hat{\mathbf{B}}_{(y; \mathbf{x}(a))} \quad \text{for } k \in s \end{aligned} \quad (6.1)$$

where

$$\begin{aligned} \hat{\mathbf{B}}_{(y; \mathbf{x})} &= \left(\sum_s a_k \mathbf{z}_k \mathbf{x}'_k\right)^{-1} \sum_s a_k \mathbf{z}_k y_k \\ &= (\hat{\mathbf{B}}'_{(y; \mathbf{x}(t))}, \hat{\mathbf{B}}'_{(y; \mathbf{x}(w))}, \hat{\mathbf{B}}'_{(y; \mathbf{x}(a))})' \end{aligned} \quad (6.2)$$

and

$$\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)} = \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k}\right)^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{k(w)} \hat{\mathbf{B}}_{(y; \mathbf{x}(w))}.$$

The term $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)}$ in the definition of \hat{e}_{1k} is the estimate of $\mathbf{B}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_U \mathbf{z}_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x}(w))}$ in (4.10). Two replacements are required in $\mathbf{B}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)}$ to arrive at $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)}$: First, sums over U are replaced by appropriately weighted sums over s_1 , giving $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)} = (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k})^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y; \mathbf{x}(w))}$. In this expression, $\mathbf{B}_{(y; \mathbf{x}(w))}$ is still unknown, so we replace it by its estimate $\hat{\mathbf{B}}_{(y; \mathbf{x}(w))}$ to arrive at $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}; \mathbf{x}_1)}$.

A key point to note is that estimates \hat{e}_{1k} can be obtained for $k \in s_1$, because $\mathbf{x}_{k(a)}$, $\mathbf{x}_{k(w)}$ and \mathbf{x}_{1k} are all known for $k \in s_1$, but estimates \hat{e}_{2k} can only be made for $k \in s$, because y_k is available only for $k \in s$. The fact that the estimates \hat{e}_{1k} are available for $k \in s_1$ rather than $k \in s$ allows us to construct (in section 7) a more efficient estimator of $V(\hat{Y}_{2Pa \text{ lin}})$ than the traditional approach to variance estimation (in section 8) where all estimated residuals are calculated only for $k \in s$.

The design weights $a_{1k} = 1/\pi_{1k}$, $a_{2k} = 1/\pi_{2k}$ and $a_k = a_{1k} a_{2k}$ were defined in section 1. In the following sections, we also need the quantities given below, defined as functions of the second-order inclusion probabilities $\pi_{1k\ell} = \Pr(k \& \ell \in s_1)$ and $\pi_{2k\ell} = \Pr(k \& \ell \in s | s_1)$:

$$a_{1k\ell} = 1/\pi_{1k\ell}, \quad a_{2k\ell} = 1/\pi_{2k\ell}, \quad a_{k\ell} = a_{1k\ell} a_{2k\ell}$$

$$D_{1k\ell} = a_{1k} a_{1\ell} - a_{1k\ell}, \quad D_{2k\ell} = a_{2k} a_{2\ell} - a_{2k\ell},$$

$$D_{k\ell} = a_k a_\ell - a_{k\ell}.$$

Here, $\pi_{2k\ell}$ and $a_{2k\ell}$ are conditional on the sample s_1 . All first-order and second-order inclusion probabilities are assumed positive. Using this notation and the above results, we now develop two different variance estimators in the next two sections.

7. The separate residual variance estimator

The variance of $\hat{Y}_{2Pa \text{ lin}}$ is given by (5.1), where e_{1k} and e_{2k} are defined by (4.10). It can be expanded as

$$V(\hat{Y}_{2P\text{ a lin}}) = V\left(\sum_{s_1} a_{1k} e_{1k}\right) + V\left(\sum_s a_k e_{2k}\right) + 2 \text{Cov}\left(\sum_{s_1} a_{1k} e_{1k}, \sum_s a_k e_{2k}\right). \quad (7.1)$$

If we knew the residuals e_{1k} and e_{2k} , unbiased estimates for these three components would be given respectively by

$$\begin{aligned} & \sum_{k \in s_1} \sum_{\ell \in s_1} D_{1k\ell} e_{1k} e_{1\ell}, \\ & \sum_{k \in s} \sum_{\ell \in s} D_{k\ell} e_{2k} e_{2\ell}, \\ & 2 \sum_{k \in s_1} \sum_{\ell \in s} D_{1k\ell} a_{2\ell} e_{1k} e_{2\ell}. \end{aligned} \quad (7.2)$$

The proof of unbiasedness is similar for all three components. For example, for the second one, we have

$$\begin{aligned} E_{s_1} E_{s|s_1} \left(\sum_{k \in s} \sum_{\ell \in s} D_{k\ell} e_{2k} e_{2\ell} \right) &= E_{s_1} \left(\sum_{k \in s_1} \sum_{\ell \in s_1} (D_{k\ell} / a_{2k\ell}) e_{2k} e_{2\ell} \right) \\ &= \sum_{k \in U} \sum_{\ell \in U} (D_{k\ell} / a_{k\ell}) e_{2k} e_{2\ell} \\ &= \sum_{k \in U} \sum_{\ell \in U} (a_k a_\ell / a_{k\ell}) e_{2k} e_{2\ell} - \left(\sum_U e_{2k} \right)^2 \\ &= E \left[\left(\sum_s a_k e_{2k} \right)^2 \right] - \left[E \left(\sum_s a_k e_{2k} \right) \right]^2 \\ &= V \left(\sum_s a_k e_{2k} \right). \end{aligned}$$

We now replace the unknown residuals in (7.2) by the respective estimates given by (6.1); that is, e_{1k} by \hat{e}_{1k} for $k \in s_1$ and e_{2k} by \hat{e}_{2k} for $k \in s$. Then, the resulting three components are added to arrive at the “separate residual” variance estimator

$$\begin{aligned} \hat{V}_{sr}(\hat{Y}_{2P\text{ a lin}}) &= \sum_{k \in s_1} \sum_{\ell \in s_1} D_{1k\ell} \hat{e}_{1k} \hat{e}_{1\ell} \\ &+ \sum_{k \in s} \sum_{\ell \in s} D_{k\ell} \hat{e}_{2k} \hat{e}_{2\ell} \\ &+ 2 \sum_{k \in s_1} \sum_{\ell \in s} D_{1k\ell} a_{2\ell} \hat{e}_{1k} \hat{e}_{2\ell}. \end{aligned} \quad (7.3)$$

The term “separate residual” and the corresponding subscript sr reflect the fact that (7.3) keeps the residuals separate, where \hat{e}_{1k} is defined over the larger sample s_1 and \hat{e}_{2k} over the smaller sample s . The fact that residuals computed for the larger sample s_1 can be advantageous for variance estimation was recognized by Axelson (1998). However, his derivation differs from our calibration approach based on \mathbf{x}_{1k} and \mathbf{x}_k . The technique for variance estimation of the two-phase regression estimator in Hidiroglou, Rao and Haziza (2006) has certain traits in

common with our approach, but there are also considerable differences.

8. The combined residual variance estimator

We arrived at (7.3) by recognizing that the estimates \hat{e}_{1k} are obtainable for $k \in s_1$. The traditional approach, reviewed in this section, is to derive a variance estimator by conditioning on the phase-one sample s_1 . This produces a variance estimator where all required residuals are defined for $k \in s$. Later, we compare it with the more efficient (7.3). From (5.1), we condition on the phase-one sample s_1 to obtain

$$\begin{aligned} V(\hat{Y}_{2P\text{ a lin}}) &= V_{s_1} E_{s|s_1} \left(\sum_{s_1} a_{1k} e_{1k} + \sum_s a_k e_{2k} \right) \\ &+ E_{s_1} V_{s|s_1} \left(\sum_{s_1} a_{1k} e_{1k} + \sum_s a_k e_{2k} \right) \\ &= V_{s_1} \left(\sum_{s_1} a_{1k} e_{1k} + \sum_{s_1} a_{1k} e_{2k} \right) \\ &+ E_{s_1} V_{s|s_1} \left(\sum_s a_k e_{2k} \right) \\ &= V_{s_1} \left(\sum_{s_1} a_{1k} e_{12k} \right) + E_{s_1} V_{s|s_1} \left(\sum_s a_k e_{2k} \right) \end{aligned} \quad (8.1)$$

where $e_{12k} = e_{1k} + e_{2k}$ is called the combined residual. From (4.10), we obtain the following.

$$\begin{aligned} e_{12k} &= y_k - \mathbf{x}'_{k(i)} \mathbf{B}_{(y;\mathbf{x})(i)} - \mathbf{x}'_k \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} \\ e_{2k} &= y_k - \mathbf{x}'_{k(i)} \mathbf{B}_{(y;\mathbf{x})(i)} - \mathbf{x}'_{k(w)} \mathbf{B}_{(y;\mathbf{x})(w)} \\ &\quad - \mathbf{x}'_{k(a)} \mathbf{B}_{(y;\mathbf{x})(a)}. \end{aligned} \quad (8.2)$$

It is straightforward to define estimators of the two components $V_{s_1}(\sum_{s_1} a_{1k} e_{12k})$ and $E_{s_1} V_{s|s_1}(\sum_s a_k e_{2k})$. Each of these has the form of a double sum over s because e_{12k} and e_{2k} contain y_k which is only available for $k \in s$. The first component uses $\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k} = y_k - \mathbf{x}'_{k(i)} \hat{\mathbf{B}}_{(y;\mathbf{x})(i)} - \mathbf{x}'_{1k} \hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\mathbf{x}_1)}$ for $k \in s$. We then have $\sum_{k \in s} \sum_{\ell \in s} D_{1k\ell} a_{2\ell} \hat{e}_{12k} \hat{e}_{12\ell}$ as an estimator of $V_{s_1}(\sum_{s_1} a_{1k} e_{12k})$.

For the second component, we use the residual estimates $\hat{e}_{2k} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}_{(y;\mathbf{x})}$ given by (6.1) for $k \in s$, and obtain $\sum_{k \in s} \sum_{\ell \in s} D_{2k\ell} a_{1k} a_{1\ell} \hat{e}_{2k} \hat{e}_{2\ell}$ as an estimator of $E_{s_1} V_{s|s_1}(\sum_s a_k e_{2k})$. Summing the two estimated terms we have the following variance estimator, where the subscript cr indicates “combined residual”,

$$\begin{aligned} \hat{V}_{cr}(\hat{Y}_{2P\text{ a lin}}) &= \sum_{k \in s} \sum_{\ell \in s} D_{1k\ell} a_{2k\ell} \hat{e}_{12k} \hat{e}_{12\ell} \\ &+ \sum_{k \in s} \sum_{\ell \in s} D_{2k\ell} a_{1k} a_{1\ell} \hat{e}_{2k} \hat{e}_{2\ell}. \end{aligned} \quad (8.3)$$

Let us review how (7.3) and (8.3) differ. The separate residual variance estimator (7.3) starts with the expansion $V(\hat{Y}_{2P\alpha lin}) = V(\sum_{s_1} a_{1k} e_{1k}) + V(\sum_s a_k e_{2k}) + 2\text{Cov}(\sum_{s_1} a_{1k} e_{1k}, \sum_s a_k e_{2k})$. We estimate these three components separately as functions of the residuals e_{1k} and e_{2k} . The resulting variance expression has three terms: a double sum over s_1 in terms of e_{1k} and e_{1j} , a double sum over s in terms of e_{2k} and e_{2j} , and a cross-sum over s_1 and s in terms of $e_{1k} \in s_1$ and $e_{2j} \in s$. Finally, we arrive at (7.3) by estimating e_{1k} by \hat{e}_{1k} for $k \in s_1$ and e_{2k} by \hat{e}_{2k} for $k \in s$.

The combined residual variance estimator (8.3) arises from the traditional conditioning on the phase-one sample s_1 as $V(\hat{Y}_{2P\alpha lin}) = V_{s_1} E_{s_1|s_1}(\hat{Y}_{2P\alpha lin}) + E_{s_1} V_{s_1|s_1}(\hat{Y}_{2P\alpha lin})$. This leads us to combine e_{1k} and e_{2k} as $e_{12k} = e_{1k} + e_{2k}$ in the first term. The second term, $E_{s_1} V_{s_1|s_1}(\hat{Y}_{2P\alpha lin})$, is a function of e_{2k} . Since e_{12k} and e_{2k} can only be estimated over s , the resulting variance estimator becomes a sum of two terms, each of them expressed as a double sum over s .

The separate residual estimator (7.3) is more efficient than the combined residual alternative (8.3), because it is based on residuals \hat{e}_{1k} obtained for the typically larger sample s_1 . The advantage of (7.3) over (8.3) is illustrated by the simulation in section 10. The approach behind the separate residual variance estimator (7.3) can be extended to three-phase sampling and other complex designs. In those extensions of the technique, we proceed in a similar manner, starting by a derivation of the linearized form through an expansion of the variance components and the determination of the appropriate residuals.

9. A comparison with the two-phase regression estimator

Särdal, Swensson and Wretman (1992) developed a two-phase regression estimator for $Y = \sum_U y_k$, based on an earlier paper by Särdal and Swensson (1989). It is useful to see how this estimator, denoted here by \hat{Y}_{reg} , compares with the calibration estimator \hat{Y}_{2P} considered in the preceding sections of this paper. When based on the same auxiliary information, the two estimators are “close” but not identical. This is because the estimator \hat{Y}_{2P} is derived by calibration in each of the two phases, whereas the two-phase regression estimator \hat{Y}_{reg} is derived by model-assisted reasoning.

We now describe the two-phase regression estimator of Särdal, Swensson and Wretman (1992). Their derivation involves the fit of two linear regression models with the use of the available auxiliary data; one at “the top level” and the other at “the bottom level”. These authors develop a corresponding estimator of variance, via the traditional conditioning argument. We compare their variance

estimator with the combined residual variance estimator (8.3), also developed by the conditioning argument. The two variance estimators do not agree exactly, because the point estimators are slightly different, but they are numerically close, as shown in this section.

Let \mathbf{x}_{1k} be a vector of auxiliary variables with known population totals, and let $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, where both \mathbf{x}_{1k} and \mathbf{x}_{2k} are known vector values for $k \in s_1$. The total $\sum_U \mathbf{x}_{1k}$ is assumed known whereas the total $\sum_U \mathbf{x}_{2k}$ is unknown. The predicted values produced for $k \in s_1$ by the two regressions fitted at the “top level” and “bottom level” are given respectively by

$$\hat{y}_{1k} = \mathbf{x}'_{1k} \hat{\mathbf{B}}_{1s} \quad \text{with} \quad (9.1)$$

$$\hat{\mathbf{B}}_{1s} = \left(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2 \right)^{-1} \left(\sum_s a_k \mathbf{x}_{1k} y_k / \sigma_{1k}^2 \right)$$

and

$$\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}_s \quad \text{with} \quad (9.2)$$

$$\hat{\mathbf{B}}_s = \left(\sum_s a_k \mathbf{x}_k \mathbf{x}'_k / \sigma_k^2 \right)^{-1} \sum_s a_k \mathbf{x}_k y_k / \sigma_k^2.$$

The resulting two-phase regression estimator \hat{Y}_{reg} of $Y = \sum_U y_k$ is

$$\begin{aligned} \hat{Y}_{reg} &= \left(\sum_U \mathbf{x}_{1k} \right)' \hat{\mathbf{B}}_{1s} + \sum_{s_1} a_{1k} (\hat{y}_k - \hat{y}_{1k}) \\ &\quad + \sum_s a_k (y_k - \hat{y}_k). \end{aligned} \quad (9.3)$$

Can \hat{Y}_{reg} be interpreted as a calibration estimator? To answer this question, let us determine the implicit weights in (9.3). We can write $\hat{Y}_{reg} = \sum_s w_k y_k$, with weights w_k identified by substituting (9.1) and (9.2) into (9.3) and simplifying. We find $w_k = a_k g_k = a_{1k} a_{2k} g_k$, where the calibration factor g_k is given for $k \in s$ by

$$\begin{aligned} g_k &= 1 + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k} \mathbf{x}_{1k} \right)' \\ &\quad \left(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2 \right)^{-1} \mathbf{x}_{1k} / \sigma_{1k}^2 \\ &\quad + \left(\sum_{s_1} a_{1k} \mathbf{x}_k - \sum_s a_k \mathbf{x}_k \right)' \\ &\quad \left(\sum_s a_k \mathbf{x}_k \mathbf{x}'_k / \sigma_k^2 \right)^{-1} \mathbf{x}_k / \sigma_k^2. \end{aligned} \quad (9.4)$$

The weights w_k are not explicitly stated in Särdal, Swensson and Wretman (1992). In what sense, if any, can w_k be considered a calibration weight? To examine this, we first replace y_k in (9.3) with \mathbf{x}'_{1k} . Using (9.1) and (9.2) with $y_k = \mathbf{x}'_{1k}$ gives $\sum_U \mathbf{x}'_{1k}$ as the right-hand side of (9.3). Thus, the weights $w_k = a_k g_k$ satisfy $\sum_s w_k \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Next we replace y_k in (9.3) with \mathbf{x}'_{2k} , again using (9.1) and (9.2) to obtain

$$\begin{aligned} \sum_{s_1} a_{1k} \mathbf{x}'_{2k} + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k} \mathbf{x}_{1k} \right)' \\ \left(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2 \right)^{-1} \\ \left(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{2k} / \sigma_{1k}^2 \right). \end{aligned} \quad (9.5)$$

Although (9.5) is an approximately unbiased estimate of the unknown \mathbf{x}_{2k} -total $\sum_U \mathbf{x}'_{2k}$, it does not have the usual form of the right-hand side of a phase-two calibration equation, such as $\sum_{s_1} a_{1k} \mathbf{x}'_{2k}$ or $\sum_{s_1} w_{1k} \mathbf{x}'_{2k}$. However, it is close. If we replace the two sums over s with appropriately weighted sums over s_1 , then (9.5) becomes $\sum_{s_1} w_{1k} \mathbf{x}'_{2k}$ where w_{1k} is given by (2.1) with $\mathbf{z}_{1k} = \mathbf{x}_{1k} / \sigma_{1k}^2$. Thus, the implicit weights w_k in \hat{Y}_{reg} calibrate exactly on the known population \mathbf{x}_{1k} -total, and they come close to calibrating on the estimated \mathbf{x}_{2k} -total $\sum_{s_1} w_{1k} \mathbf{x}'_{2k}$. This suggests that \hat{Y}_{reg} should have properties similar to an estimator \hat{Y}_{2P} obtained by defining \mathbf{x}_k in \hat{Y}_{2P} as $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ with $\mathbf{x}_{k(l)} = \mathbf{x}_{1k}$, $\mathbf{x}_{k(w)} = \mathbf{x}_{2k}$ and $\mathbf{x}_{k(a)} = \phi$. In addition, the form of the model-assisted estimator implies $\mathbf{z}_{1k} = \mathbf{x}_{1k} / \sigma_{1k}^2$ and $\mathbf{z}_k = \mathbf{x}_k / \sigma_k^2$. Since \mathbf{x}_k includes \mathbf{x}_{1k} it is reasonable to define $\mathbf{z}_k = \mathbf{x}_k / \sigma_k^2$ as $\mathbf{z}_k = (\mathbf{x}'_{1k} / \sigma_{1k}^2, \mathbf{x}'_{2k} / \sigma_{2k}^2)'$. These specifications meet the requirements for asymptotic equivalence of \hat{Y}_{2Pa} and \hat{Y}_{2Pw} so we do not need to worry about the choice of starting weights in \hat{Y}_{2P} . We can simply work with \hat{Y}_{2Pa} as the estimator comparable to \hat{Y}_{reg} . Now, let us look at variance estimation for \hat{Y}_{reg} and the estimator \hat{Y}_{2Pa} under these specifications.

The variance estimator of Särndal, Swensson and Wretman (1992) contains calibration factors denoted g_{ks} and g_{1ks_1} . They are not to be confused with g_k given by (9.4). If we disregard g_{ks} and g_{1ks_1} , both of which are near one and of limited numerical impact, their variance estimator is

$$\begin{aligned} \hat{V}(\hat{Y}_{\text{reg}}) = \sum_{k \in s} \sum_{f \in s} D_{1kf} a_{2kf} \hat{e}_{1ks} \hat{e}_{1fs} \\ + \sum_{k \in s} \sum_{f \in s} D_{2kf} a_{1kf} \hat{e}_{ks} \hat{e}_{fs} \end{aligned} \quad (9.6)$$

where, for $k \in s$,

$$\hat{e}_{1ks} = y_k - \mathbf{x}'_{1k} \hat{\mathbf{B}}_{1s} \text{ and } \hat{e}_{ks} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}_s. \quad (9.7)$$

Both components of (9.6) are double sums over s , reflecting the fact that both \hat{e}_{1ks} and \hat{e}_{ks} can only be obtained for $k \in s$. Formula (9.6) looks similar to formula (8.3) for the combined residual estimator but how different are the residuals in the two formulas? Let us look at the residuals for the comparable point estimator. As noted above, this estimator \hat{Y}_{2P} has $\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ with $\mathbf{x}_{k(l)} = \mathbf{x}_{1k}$, $\mathbf{x}_{k(w)} = \mathbf{x}_{2k}$, $\mathbf{x}_{k(a)} = \phi$, $\mathbf{z}_{1k} = \mathbf{x}_{1k} / \sigma_{1k}^2$ and $\mathbf{z}_k = \mathbf{x}_k / \sigma_k^2 = (\mathbf{x}'_{1k} / \sigma_{1k}^2, \mathbf{x}'_{2k} / \sigma_{2k}^2)'$. Under these specifications, the residuals \hat{e}_{1k} and \hat{e}_{2k} in (6.1) are given by

$$\begin{aligned} \hat{e}_{1k} &= \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))} - \mathbf{x}'_k \hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)}; \mathbf{x}_1)} \quad \text{for } k \in s_1 \\ \hat{e}_{2k} &= y_k - \mathbf{x}'_k \hat{\mathbf{B}}_{(y; \mathbf{x})} \\ &= y_k - \mathbf{x}'_{1k} \hat{\mathbf{B}}_{(y; \mathbf{x}(1))} - \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))} \quad \text{for } k \in s \end{aligned} \quad (9.8)$$

where $\hat{\mathbf{B}}_{(y; \mathbf{x})} = (\hat{\mathbf{B}}'_{(y; \mathbf{x}(1))}, \hat{\mathbf{B}}'_{(y; \mathbf{x}(2))})'$ corresponds to the partitioning of $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ and from (6.2)

$$\begin{aligned} \hat{\mathbf{B}}_{(y; \mathbf{x})} &= \left(\sum_s a_k \mathbf{z}_k \mathbf{x}'_k \right)^{-1} \left(\sum_s a_k \mathbf{z}_k y_k \right) \\ \hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)}; \mathbf{x}_1)} &= \left(\sum_{s_1} a_{1k} \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2 \right)^{-1} \\ &\quad \left(\sum_{s_1} a_{1k} \mathbf{x}_{1k} \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))} / \sigma_{1k}^2 \right). \end{aligned} \quad (9.9)$$

The residuals \hat{e}_{2k} in (9.8) are the same as \hat{e}_{ks} in (9.7). But how do the residuals $\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k}$, obtained by adding in (9.8), relate to their counterparts \hat{e}_{1ks} in (9.7)? To find this link, we first show that $\hat{\mathbf{B}}_{1s} = (\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2)^{-1} \sum_s a_k \mathbf{x}_{1k} y_k / \sigma_{1k}^2$ can be written as

$$\begin{aligned} \hat{\mathbf{B}}_{1s} &= \hat{\mathbf{B}}_{(y; \mathbf{x}(1))} \\ &+ \left(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2 \right)^{-1} \left(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))} / \sigma_{1k}^2 \right). \end{aligned} \quad (9.10)$$

To see this, we start with $\hat{\mathbf{B}}_{(y; \mathbf{x})}$, which by definition satisfies $\sum_s a_k \mathbf{z}_k y_k = (\sum_s a_k \mathbf{z}_k \mathbf{x}'_k) \hat{\mathbf{B}}_{(y; \mathbf{x})}$. This equality can also be written as $\sum_s a_k \mathbf{z}_k y_k = \sum_s a_k \mathbf{z}_k (\mathbf{x}'_{1k} \hat{\mathbf{B}}_{(y; \mathbf{x}(1))} + \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))})$. Since $\mathbf{z}_k = (\mathbf{x}'_{1k} / \sigma_{1k}^2, \mathbf{x}'_{2k} / \sigma_{2k}^2)'$, the component of this equation corresponding to \mathbf{x}_{1k} is $\sum_s a_k \mathbf{x}_{1k} y_k / \sigma_{1k}^2 = \sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} \hat{\mathbf{B}}_{(y; \mathbf{x}(1))} / \sigma_{1k}^2 + \sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))} / \sigma_{1k}^2$. Premultiplying both sides by $(\sum_s a_k \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2)^{-1}$, we obtain (9.10).

Then, starting with (9.8) and using the definition of $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)}; \mathbf{x}_1)}$ given by (9.9), we have

$$\begin{aligned} \hat{e}_{12k} &= \hat{e}_{1k} + \hat{e}_{2k} \\ &= y_k - \mathbf{x}'_{1k} \hat{\mathbf{B}}_{(y; \mathbf{x}(1))} - \mathbf{x}'_k \hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)}; \mathbf{x}_1)} \\ &= y_k - \mathbf{x}'_{1k} \left\{ \hat{\mathbf{B}}_{(y; \mathbf{x}(1))} + \left(\sum_{s_1} a_{1k} \mathbf{x}_{1k} \mathbf{x}'_{1k} / \sigma_{1k}^2 \right)^{-1} \right. \\ &\quad \left. \left(\sum_{s_1} a_{1k} \mathbf{x}_{1k} \mathbf{x}'_{2k} \hat{\mathbf{B}}_{(y; \mathbf{x}(2))} / \sigma_{1k}^2 \right) \right\}. \end{aligned}$$

In the expression within curly brackets, let us replace the two a_{1k} -weighted sums over s_1 with the corresponding a_k -weighted sums over s ; the result is equal to $\hat{\mathbf{B}}_{1s}$ as given by (9.10). This means $\hat{e}_{12k} \cong y_k - \mathbf{x}'_{1k} \hat{\mathbf{B}}_{1s} = \hat{e}_{1ks}$. In summary, $\hat{e}_{12k} \cong \hat{e}_{1ks}$ for $k \in s$ and $\hat{e}_{2k} = \hat{e}_{ks}$ for $k \in s$. Hence, the variance estimator (9.6) for the two-phase regression estimator \hat{Y}_{reg} should be numerically close to the combined residual variance estimator (8.3) for the calibration estimator \hat{Y}_{2P} defined in this section. We present empirical support for this through the simulation in next section.

10. Simulation

In this section we present a small simulation to validate the claim that the separate residual variance estimator $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$ given by (7.3) can be considerably more efficient than the combined residual variance estimator $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$ given by (8.3), and that the behaviour of the latter is very similar to that of the two-phase regression estimator $\hat{V}(\hat{Y}_{\text{reg}})$ given by (9.6). We created a population of $N = 5,000$ units in two steps as follows: First, the values (u_{1k}, u_{2k}) for $k = 1, 2, \dots, 5,000$ were generated by 5,000 realizations of the independent random variables $u_{1k} \sim 2\text{Gamma}(4)$ and $u_{2k} \sim 3\text{Gamma}(6)$, where the $\text{Gamma}(a)$ distribution has density $f(x) = [\Gamma(a)]^{-1} x^{a-1} e^{-x}$ for $x > 0$. Secondly, the values of the variable of interest were created as $y_k = 10 + u_{1k} + 3u_{2k} + \varepsilon_k$, $k = 1, 2, \dots, 5,000$, with $\varepsilon_k \sim 5\text{Normal}(0)$, where $\text{Normal}(0)$ is the standard Normal distribution with mean 0 and variance 1. The target of estimation in the experiment is the population y -total $Y = \sum U y_k = 358,205$. For the phase-one calibration, we used the auxiliary vector $\mathbf{x}_{1k} = (1, u_{1k})'$ and $\mathbf{z}_{1k} = \mathbf{x}_{1k}$. That is, the weights w_{1k} for $k \in s_1$ were determined by calibration to the known total $(N, \sum_U u_{1k}) = (5,000, 39,611.8)$. For the phase-two calibration we used $\mathbf{x}_k = (\mathbf{x}_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$ with $\mathbf{x}_{k(t)} = (1, u_{1k})'$, $\mathbf{x}_{k(w)} = u_{2k}$, $\mathbf{x}_{k(a)} = \phi$ and $\mathbf{z}_k = \mathbf{x}_k$. These specifications satisfy the conditions for asymptotic equivalence between \hat{Y}_{2Pa} and \hat{Y}_{2Pw} . Therefore, for this simulation, we can work with \hat{Y}_{2Pa} and its linearized form $\hat{Y}_{2Pa\text{lin}}$.

For each phase-one sample s_1 , the final weights w_k for the estimator $\hat{Y}_{2Pa} = \sum s w_k y_k$ were determined by calibrating to the known totals given by the vector $(N, \sum_U u_{1k}, \sum_{s_1} w_{1k} u_{2k}) = (5,000, 39,611.8, \sum_{s_1} w_{1k} u_{2k})$. It is important to note that it was not necessary to have $\mathbf{x}_{k(a)} = \phi$ in order to run a simulation to compare

$\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$ and $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$. However, we can not compare $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$ and $\hat{V}(\hat{Y}_{\text{reg}})$ unless we define an estimator \hat{Y}_{2Pa} comparable to \hat{Y}_{reg} , and to achieve this we need $\mathbf{x}_{k(a)} = \phi$, as noted in section 9.

We drew repeated sample pairs (s_1, s) , where s_1 is an SRS of n_1 units from U , and s is an SRS of n units from s_1 . Here SRS stands for simple random sampling without replacement. We worked with different size combinations (n_1, n) : (4000, 3000), (4000, 2000), (4000, 1000), (3000, 2000), (3000, 1000) and (2000, 1000). If $n = n_1$, two-phase sampling is equivalent to one-phase sampling, and $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$ and $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$ are identical.

For each combination (n_1, n) , we realized 100,000 sample pairs (s_1, s) . Based on the data for each of these outcomes, we computed the separate residual variance estimator $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$, the combined residual variance estimator $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$ and the variance estimator $\hat{V}(\hat{Y}_{\text{reg}})$. For this purpose, we used the respective expressions that follow from (7.3), (8.3) and (9.6) when SRS is specified at each phase. To save space, these expressions are not shown here. We obtained 100,000 realized values for each of the three variance estimators. Figure 10.1 shows the distributions of the 100,000 \hat{V} -values for $n_1 = 4,000$ and $n = 2,000$.

The figure shows strikingly different distributions for $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$ and $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$. The distribution of the separate residual estimator $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$ is much more concentrated. Thus $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$ is more efficient than $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$ and on average, it produces considerably shorter confidence intervals. We also note that the distribution of $\hat{V}(\hat{Y}_{\text{reg}})$ is very similar to that of $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$. This supports our analysis in section 9. Similar results were obtained for the other sample sizes in the simulation.

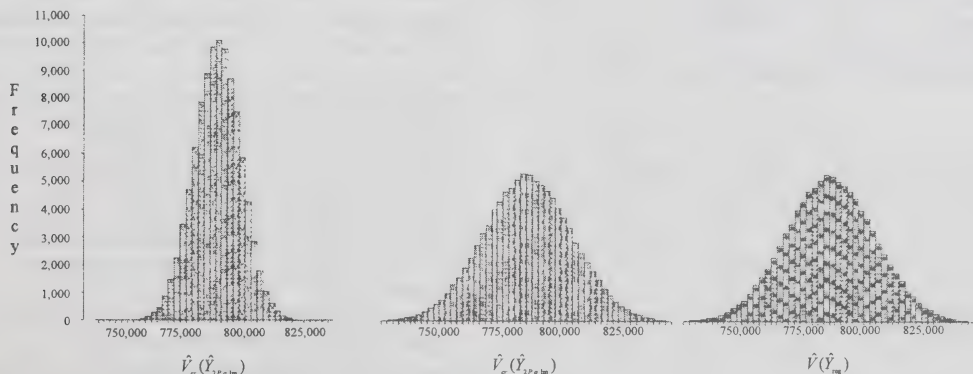


Figure 10.1 Distribution of 100,000 realized values for $\hat{V}_{sr}(\hat{Y}_{2Pa\text{lin}})$, $\hat{V}_{cr}(\hat{Y}_{2Pa\text{lin}})$ and $\hat{V}(\hat{Y}_{\text{reg}})$

To obtain a measure of the efficiency of the three variance estimators, we computed the simulation variance of the 100,000 \hat{V} -values. These simulation variances are shown in Table 10.1, Table 10.2 and Table 10.3. The numbers are dramatically lower for $\hat{V}_{sr}(\hat{Y}_{2Pa lin})$ than for the other two. Table 10.4 shows the relative advantage of $\hat{V}_{sr}(\hat{Y}_{2Pa lin})$ over $\hat{V}_{cr}(\hat{Y}_{2Pa lin})$. For this population, the simulation variance of $\hat{V}_{sr}(\hat{Y}_{2Pa lin})$ is less than half the simulation variance of $\hat{V}_{cr}(\hat{Y}_{2Pa lin})$.

Table 10.1
Simulation variance for the separate residual variance estimator $\hat{V}_{sr}(\hat{Y}_{2Pa lin})$

n_1	n		
	3,000	2,000	1,000
4,000	64.82	95.91	484.92
3,000		1,179.62	1,806.79
2,000			13,995.94

Note: Actual values are the displayed values times 10^6 .

Table 10.2
Simulation variance for the combined residual variance estimator $\hat{V}_{cr}(\hat{Y}_{2Pa lin})$

n_1	n		
	3,000	2,000	1,000
4,000	153.22	364.08	1,290.41
3,000		2,449.05	6,855.69
2,000			33,220.88

Note: Actual values are the displayed values times 10^6 .

Table 10.3
Simulation variance for the variance estimator $\hat{V}(\hat{Y}_{reg})$

n_1	n		
	3,000	2,000	1,000
4,000	153.25	364.14	1,289.79
3,000		2,449.36	6,854.52
2,000			33,210.31

Note: Actual values are the displayed values times 10^6 .

Table 10.4
Ratio of entries in Table 10.1 to corresponding entries in Table 10.2

n_1	n		
	3,000	2,000	1,000
4,000	0.42	0.26	0.38
3,000		0.48	0.26
2,000			0.42

11. Discussion

In a design-based perspective on estimation for two-phase sampling designs, one can follow a regression estimation approach or a calibration estimation approach. We concentrate on the calibration approach to create approximately design-unbiased estimators. The extent of the information available for the calibration holds the key to the efficiency of the estimates. We recognize in this paper that there are three different types of auxiliary variables associated with two-phase designs. They have different information characteristics. From these we define four different auxiliary vectors; one for the phase-one calibration and the other three for the phase-two calibration. The calibration approach is suitable for analyzing the resulting estimators in a systematic manner. As the paper shows, this approach also leads to a more efficient variance estimator than the traditional method for variance estimation in two-phase designs.

References

Axelsson, M. (1998). Variance estimation for the generalised regression estimator under two-phase sampling - a modified approach. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 85-89.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Dupont, F. (1995). Alternative adjustments when there are several levels of auxiliary information. *Survey Methodology*, 21, 125-136.

Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two phase sampling. *Journal of Official Statistics*, 18, 233-255.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143-154.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

Hidiroglou, M.A., Rao, J.N.K. and Haziza, D. (2006). Variance estimation in two phase sampling. (Accepted paper to appear in) *Australian and New Zealand Journal of Statistics*.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-90.

Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Survey weighted hat matrix and leverages

Jianzhu Li and Richard Valliant¹

Abstract

Regression diagnostics are geared toward identifying individual points or groups of points that have an important influence on a fitted model. When fitting a model with survey data, the sources of influence are the response variable Y , the predictor variables X , and the survey weights, W . This article discusses the use of the hat matrix and leverages to identify points that may be influential in fitting linear models due to large weights or values of predictors. We also contrast findings that an analyst will obtain if ordinary least squares is used rather than survey weighted least squares to determine which points are influential.

Key Words: Influence; Linear regression; Survey data; Weighted least squares.

1. Introduction

In some conventional linear regression diagnostics, it is often useful to measure the influence each data point can have in determining the values of parameter estimates and, in turn, fitted values. The hat matrix and its diagonal elements, referred to as leverages, are popular techniques that are used to identify the cases that have outlying values for predictor variables, and, therefore, may be influential in model fitting if they are also associated with unusual residuals. When there is more than one predictor variable in the regression, analysts can compute leverages to summarize the collective influence of the X values for each observation.

In finite population estimation, a superpopulation assumption is usually used to build models. Suppose that some model fits reasonably well for the bulk of the population. For convenience, we will refer to this as the “true” model. However, the goal is usually to find a model that has some descriptive or predictive power, bearing in mind that no model is really “true”. The influence diagnostics should allow analysts to identify points that make estimated parameters deviate from that true model. Parameter estimates in linear regression using complex survey data are often derived from the pseudo maximum likelihood approach, outlined by Skinner, Holt and Smith (1989, Chapter 3), following ideas of Binder (1983). In this paper, we assume that the analyst has decided that an estimator involving sample weights is appropriate for his or her problem. As shown in later sections, the survey weighted hat matrix and leverages are useful for detecting potentially influential observations caused by not only extreme X values, but also by large sample weights.

Previous survey literature has discussed the effect of outliers on some survey estimates, but does not give much attention to diagnostics for linear regression models. Deville

and Särndal (1992), and Potter (1990, 1993) discuss some possibilities for locating or trimming extreme survey weights when the goal is to estimate population totals and other simple descriptive statistics. Hulliger (1995) and Moreno-Rebollo, Muñoz-Reyes and Muñoz-Pichardo (1999) address the effect of outliers on the Horvitz-Thompson estimator of a population total. Smith (1987) demonstrates diagnostics based on case deletion and a form of the influence function. Zaslavsky, Schenker and Belin (2001), and Beaumont and Alavi (2004) use M-estimation based strategies to downweight the influential clusters or units. Chambers (1986), Gwet and Rivest (1992), Welsh and Ronchetti (1998), and Duchesne (1999) conduct research on outlier robust estimation techniques for totals.

A perennial question among analysts of survey data is whether to use the survey weights or not when fitting models. The collections edited by Skinner *et al.* (1989) and Chambers and Skinner (2003) discuss this issue at length. Binder and Roberts (2003, Chapter 3), Chambers, Dorfman and Sverchkov (2003, Sections 11.2.3, 11.6), Chambers and Skinner (2003, Chapter 1), Korn and Graubard (1999, Sections 4.3, 4.4), Pfeffermann (1996), and Smith (1989, Chapter 6) describe the arguments pro and con. The details can be quite mathematical and abstract but are summarized succinctly by Skinner (2003, Section 6.2.3).

We paraphrase Skinner (2003, Section 6.2.3) here in the context of fitting a linear model to predict some response Y based on a set of explanatory variables X . If the linear model is specified correctly and the sampling depends only on the explanatory variables in the model, then unweighted regression parameter estimates will be unbiased in a model-based sense. In particular, the assumed conditions require that the survey weights are unrelated to Y conditional on the values of the X predictors. However, if sampling depends on factors that may be related to Y , even after conditioning on the values of the predictors, the unweighted parameter

1. Jianzhu Li, Westat, 1650 Research Boulevard, Rockville MD 20850; Richard Valliant, Survey Research Center, University of Michigan, and Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742

estimators will be biased both with respect to the true model and in the design-based, repeated sampling sense. This situation is known as having an *informative* sample design in which the distribution of the sample values of Y is different from the population distribution. An example of this is given by Chambers, Dorfman and Sverchkov (2003, Section 11.2.3). If sample units are selected with probabilities proportional to some measure x of their size and Y is related to x , the sample distribution of Y will be skewed to the right of its population distribution. The situation in this example is similar to the one in our empirical study in section 5.

Using the survey weights guards against the bias that may result from not accounting for an informative sample. Also, if the model is not correctly specified, the survey-weighted regression still estimates a census parameter. That is, the weighted estimates are approximately unbiased for the best-fitting linear model that would be obtained if the entire finite population were in hand. In this paper, we assume that an analyst has made the decision to use weights in fitting a model, possibly for the reasons above, and provide one type of diagnostic for assessing the effects of certain data points.

The hat matrix and leverages we present are the same ones that are produced by standard software packages when a weighted least squares regression is done. However, the literature is missing any discussion of their use and interpretation in the context of survey-weighted regression. Korn and Graubard (1999) is one of the few references that addresses any kind of diagnostics for models fitted from survey data. Leverages are among a series of diagnostic tools and will be more effective when evaluated with residuals. Many diagnostic statistics, such as the famous Cook's distance (Cook 1977) turn out to have both leverages and residuals as components.

The literature gives somewhat ambiguous guidance on how to deal with the influential observations once they are identified. An obvious, and perhaps naïve, solution is to remove the outliers and refit the model, which makes sense when the outliers result from improperly recorded data. A natural extension of this would be to devise an automatic approach where certain rules would be used to identify influential points, delete them, and refit the model. Our presumption in this article is that, after identification of influential points and careful consideration of the reasons for the influence, an analyst will determine whether the points should be excluded from fitting. This is in contrast to setting up some procedure that would automatically exclude points based on some cutoff values.

The remainder of the paper is organized as follows. Section 2 describes the ordinary least squares hat matrix, leverages, and some of their properties. Sections 3 and 4

cover the survey-weighted hat matrix and leverages plus a decomposition that shows how points can have large leverages. The extensions to survey data apply to both single- and multi-stage designs. Section 5 gives a numerical example using a single-stage sample of mental health organizations. The last section summarizes our findings and gives some directions for additional research.

2. OLS hat matrix

A *working* model is one that is being provisionally considered by an analyst for the structure that best describes a conceptual superpopulation. It may be revised after further assessment by adding predictors, dropping predictors, or making other changes to the form of the model. Suppose that the working linear model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} \quad (1)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{X}^T = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$. Assuming the \mathbf{X} matrix is of full rank, the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{A}^{-1} \mathbf{X}^T \mathbf{Y}, \quad (2)$$

where $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ is a square matrix and invertible. The fitted values $\hat{\mathbf{Y}}$ corresponding to the observed values \mathbf{Y} are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T$ is called the hat matrix. This name was first introduced by Tukey (Belsley, Kuh and Welsch 1980, Chapter 2; Hoaglin and Welsch 1978). The leverage, $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i$, is the i^{th} element on the diagonal of the hat matrix, which measures the impact of Y_i on its own fitted value since $\hat{Y}_i = \sum_j h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$. If h_{ii} approaches 1, Y_i has a crucial role in determining the value of \hat{Y}_i .

The OLS hat matrix and leverages have many special and useful properties:

- (i) \mathbf{H} is symmetric, or $h_{ij} = h_{ji}$;
- (ii) \mathbf{H} is idempotent, or $\mathbf{H} = \mathbf{H}^2$, or $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$;
- (iii) $\mathbf{H}\mathbf{X} = \mathbf{X}$ or $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$;
- (iv) $0 \leq h_{ii} \leq 1$;
- (v) $\sum_i h_{ii} = \text{rank}(\mathbf{X}) = p$, which implies that the mean leverage is $\bar{h} = p/n$;

if model (1) has an intercept, the following two properties hold:

- (vi) $\sum_i h_{ij} = 1$;

$$(vii) h_{ii} = 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{A}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \text{ where } \bar{\mathbf{x}} = \sum_n \mathbf{x}_i / n.$$

In a reasonably large data set, an individual leverage value h_{ii} is usually considered extreme if it is more than twice the mean, $\bar{h} = p/n$ (Belsley *et al.* 1980, Chapter 2). The existence of a gap between most of the cases and a few unusual cases in the empirical distribution of the leverages also provides evidence of outlying units.

3. Survey weighted hat matrix

The initial step in the pseudo maximum likelihood approach is to form the set of estimating equations that would be appropriate for a model if the entire finite population were observed. This set is a type of population total which is then estimated using design-based survey methods. Suppose that the underlying structural model is a fixed-effects linear model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind } N(0, v_i \sigma^2) \quad (3)$$

where ε_i is independently normally distributed with mean 0 and variance $v_i \sigma^2$, which is known except for the constant σ^2 . The pseudo maximum likelihood estimator (PMLE) of $\boldsymbol{\beta}$ is the solution to the set of estimating equations $\mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}) = 0$, with $\mathbf{V} = \text{diag}(v_1, \dots, v_n)$ and $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$. Survey weights, which in probability samples are usually inversely proportional to inclusion probabilities, are used in the PMLE to account for an informative design in which the sample distribution of the Y 's is likely to differ from that of the finite population. These equations can be solved explicitly as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{Y}$. If we assume $\mathbf{V} = \mathbf{I}$, model (3) reduces to (1) and the survey-weighted (SW) estimator $\hat{\boldsymbol{\beta}}$ will consequently take the form of a weighted least squares estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$.

When survey weights are accounted for in the regression, the predicted values become $\hat{\mathbf{Y}} = \mathbf{H} \mathbf{Y}$, where the hat matrix includes the survey weights and is defined as

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} = \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W}$$

with $\mathbf{A} = \mathbf{X}^T \mathbf{W} \mathbf{X}$. The leverages on the diagonal of the hat matrix are $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i w_i$. In this formulation, it is assumed that the analyst does not incorporate a \mathbf{V} matrix in the regression. However, results can be modified to incorporate \mathbf{V} simply by using $\mathbf{W}^* = \mathbf{W} \mathbf{V}^{-1}$ rather than \mathbf{W} . Unlike the unweighted hat matrix, the SW hat matrix is no longer symmetric for sampling designs with unequal selection probabilities (or, more generally, unequal weights). Properties (ii) – (vi) in section 2 still hold (*e.g.*, see Valliant, Dorfman and Royall 2000, Chapter 5) provided the unweighted hat matrices were replaced by the weighted

ones. In addition, the SW hat matrix has extra useful, and easily verified, properties as follows:

- a) $\mathbf{W} \mathbf{H} = \mathbf{W} \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} = \mathbf{H}^T \mathbf{W}$;
- b) $\mathbf{X}^T \mathbf{W} (\mathbf{I} - \mathbf{H}) = \mathbf{X}^T \mathbf{W} - \mathbf{X}^T \mathbf{H}^T \mathbf{W} = \mathbf{0}$;
- c) $w_i h_{ii} = w_i \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i w_i = w_i h_{ii}$.

The definition of the weighted leverages indicates that a large leverage may be caused by outlying \mathbf{X} values, an outlying weight, or both. Note that the formulas for the survey-weighted hat matrix and leverages apply regardless of whether the sample design uses strata or is single-stage or multi-stage. This is in contrast to diagnostics, like Cook's D , that require estimated standard errors or covariance matrices that should be specialized to fit the sample design.

4. Decomposition of leverages

Leverages can be decomposed into components that separate the effect of the weight and the \mathbf{X} values for a unit. Suppose the working model is (1) and that the model contains an intercept, so that

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^T \end{pmatrix} \equiv (\mathbf{1} \mid \mathbf{X}_1), \text{ and } \mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

where $\mathbf{x}_i^T = (x_{i1}, \dots, x_{i,p-1})$ are $1 \times (p-1)$ vectors, $\mathbf{1}$ is a $n \times 1$ vector with all the elements equal to 1, and \mathbf{X}_1 is a $n \times (p-1)$ matrix. The \mathbf{A} matrix is computed as

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{X}_1^T \end{pmatrix} \mathbf{W} (\mathbf{1} \mid \mathbf{X}_1) = \begin{pmatrix} \mathbf{1}^T \mathbf{W} \mathbf{1} & \mathbf{1}^T \mathbf{W} \mathbf{X}_1 \\ \mathbf{X}_1^T \mathbf{W} \mathbf{1} & \mathbf{X}_1^T \mathbf{W} \mathbf{X}_1 \end{pmatrix} \equiv \begin{pmatrix} \hat{N} & \hat{\mathbf{t}}_X^T \\ \hat{\mathbf{t}}_X & \mathbf{A}_1 \end{pmatrix},$$

where $\hat{\mathbf{t}}_X$ is a $(p-1) \times 1$ vector with elements $\hat{t}_{x_j} = \sum_{i \in s} w_i x_{ij}$ and \mathbf{A}_1 is a $(p-1) \times (p-1)$ matrix. Using the inverse of a partitioned matrix,

$$\begin{aligned} \mathbf{A}^{-1} &= \begin{pmatrix} \frac{1}{\hat{N}} + \frac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \hat{\mathbf{t}}_X & -\frac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \\ -\frac{1}{\hat{N}} \mathbf{S}^{-1} \hat{\mathbf{t}}_X & \mathbf{S}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\hat{N}} + \bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \bar{\mathbf{x}}_W & -\bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \\ -\mathbf{S}^{-1} \bar{\mathbf{x}}_W & \mathbf{S}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\hat{N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\bar{\mathbf{x}}_W^T \\ \mathbf{I} \end{pmatrix} \mathbf{S}^{-1} (-\bar{\mathbf{x}}_W \mid \mathbf{I}) \end{aligned}$$

where $\bar{\mathbf{x}}_w = \hat{\mathbf{t}}_x / \hat{N}$ is a $(p-1) \times 1$ vector, and $\mathbf{S} = \mathbf{A}_1 - \hat{\mathbf{t}}_x \hat{\mathbf{t}}_x^T / \hat{N}$ is a $(p-1) \times (p-1)$ matrix. Simplifying the hat matrix using the above inverse matrix, we obtain

$$\begin{aligned} \mathbf{H} &= \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} \\ &= \left\{ \frac{1}{\hat{N}} \mathbf{1} \mathbf{1}^T + (\mathbf{X}_1 - \mathbf{1} \bar{\mathbf{x}}_w^T) \mathbf{S}^{-1} (-\bar{\mathbf{x}}_w \mathbf{1}^T + \mathbf{X}_1^T) \right\} \mathbf{W} \\ &= \left\{ \frac{1}{\hat{N}} \mathbf{1} \mathbf{1}^T + \begin{pmatrix} \mathbf{x}_1^T - \bar{\mathbf{x}}_w^T \\ \vdots \\ \mathbf{x}_n^T - \bar{\mathbf{x}}_w^T \end{pmatrix} \mathbf{S}^{-1} (\mathbf{x}_1 - \bar{\mathbf{x}}_w, \dots, \mathbf{x}_n - \bar{\mathbf{x}}_w) \right\} \mathbf{W}. \end{aligned}$$

Then, using the fact that $\hat{N} = n\bar{w}$ with $\bar{w} = \sum_{i=1}^n w_i / n$, the leverage of i^{th} observation, or the i^{th} diagonal element of the weighted hat matrix \mathbf{H} , is

$$h_{ii} = \frac{1}{n} \frac{w_i}{\bar{w}} [1 + \hat{N} (\mathbf{x}_i - \bar{\mathbf{x}}_w)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_w)].$$

The quadratic form, $(\mathbf{x}_i - \bar{\mathbf{x}}_w)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_w)$, defines an ellipsoid centered at $\bar{\mathbf{x}}_w$ (e.g., see Weisberg 2005, Chapter 8), and $\hat{N} (\mathbf{x}_i - \bar{\mathbf{x}}_w)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_w)$ is the Mahalanobis distance from \mathbf{x}_i to $\bar{\mathbf{x}}_w$. Consequently, a leverage can be large if (1) w_i is large, especially relative to the average weight \bar{w} ; or (2) \mathbf{x}_i is far from the weighted average, $\bar{\mathbf{x}}_w$, of the \mathbf{X} , in the metric determined by the matrix \mathbf{S} .

For example, in a simple linear model with only one auxiliary variable, $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2)$, the leverage of the i^{th} observation is

$$h_{ii}^w = \frac{1}{n} \frac{w_i}{\bar{w}} \left[1 + \hat{N} \frac{(x_i - \bar{x}_w)^2}{\sum_{j=1}^n w_j (x_j - \bar{x}_w)^2} \right].$$

where $\bar{x}_w = \sum_i w_i x_i / \hat{N}$.

If the error terms in the model have a general variance structure $\varepsilon \sim (0, \mathbf{V})$ and \mathbf{V} is known, the hat matrix is then defined as $\mathbf{H} = \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V}^{-1}$ with

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} \mathbf{1}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{1} & \mathbf{1}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X}_1 \\ \mathbf{X}_1^T \mathbf{V}^{-1} \mathbf{W} \mathbf{1} & \mathbf{X}_1^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X}_1 \end{pmatrix} \\ &= \begin{pmatrix} \sum_s w_s / v_s & \sum_s w_s \mathbf{x}_s^T / v_s \\ \sum_s w_s \mathbf{x}_s / v_s & \sum_s w_s \mathbf{x}_s \mathbf{x}_s^T / v_s \end{pmatrix}. \end{aligned}$$

A formula for \mathbf{A}^{-1} like the one above applies with $\hat{\mathbf{t}}_{xV} = \sum_s w_s \mathbf{x}_s / v_s$, $\hat{N}_V = \sum_s w_s / v_s$, and $\mathbf{S}_V = \mathbf{X}_1^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X}_1 - \hat{\mathbf{t}}_{xV} \hat{\mathbf{t}}_{xV}^T / \hat{N}_V$. If a general \mathbf{V} is used, $\hat{\mathbf{t}}_{xV}$ and \hat{N}_V no longer are design-based estimates of \mathbf{T}_x and N but are estimates of $\mathbf{T}_{xV} = \sum_i^N \mathbf{x}_i / v_i$ and $N_V = \sum_i^N 1 / v_i$. The leverage of the i^{th} observation under this general model is

$$h_{ii} = \frac{w_i}{v_i \hat{N}_V} [1 + \hat{N}_V (\mathbf{x}_i - \bar{\mathbf{x}}_{wV})^T \mathbf{S}_V^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{wV})].$$

5. Numerical example

As noted in section 1, arguments can be advanced to justify ignoring sample design features, generally, and weights, in particular, when fitting models. Roughly speaking, when a model conditions on all the design variables determining the sampling scheme and the model is correct for both the population and the sample, OLS regression can be used. Analysts may object to including design variables in a model because some are not scientifically interesting as predictors. In addition, conditioning on all design variables may not be possible, especially when the “sampling scheme” includes uncontrolled nonresponse that itself may be related to the response variable. As noted in section 1, SW provides a modicum of protection against having a misspecified model when the distribution of the sample Y 's is different from that of the population due to the type of sample design used. Nevertheless, some analysts will contend that the sample design and survey weights can be ignored in specific applications and that OLS is appropriate. Thus, it is interesting to see how different the OLS diagnostics are from SW diagnostics in a real application. However, given a course of action, an analyst should use diagnostics consistent with the method of fitting. If OLS is used, the standard OLS diagnostics should be examined; if SW regression is used, SW diagnostics are appropriate. It may well be that different points are influential depending on whether one uses OLS or SW regression.

In this section we examine the hat matrix and leverages in a regression example using the 1998 Survey of Mental Health Organizations (SMHO) conducted in the U.S., which collected data on specialty mental health care organizations and general hospital mental health care services. The sample for this survey was based on a stratified single-stage design with probability proportional to size (PPS) sampling (Manderscheid and Henderson 2002; Choudhry 2000). The measure of size (MOS) used in sampling was the number of “episodes”, defined as the number of patients/clients of an organization at the beginning of 1998 plus the number of new patients/clients added during calendar year 1998. Many of the analysis variables in the survey are related to the MOS, and their unweighted sample distributions will be different from the population distributions since the sample tends to have larger size units. Thus, this design is potentially informative as defined in Chambers and Skinner (2003).

The varying sizes of the mental health care organizations resulted in the values of collected variables in the sample having wide ranges, which may cause some observations to have relatively large influence on the parameter estimates of a linear regression. The model of interest in this study is to regress the total expenditure of a health organization, in 1,000's of dollars, on the number of beds set up and staffed for use and the number of additions of patients or clients during the reporting year. The SW estimator, $\hat{\beta} = (X^T W X)^{-1} X^T W Y$, was used. Mimicking the procedure employed by most analysts, we did not incorporate a model variance matrix **V** in the estimate of the regression parameter. A total of 875 observations was used in the regression, each of which had non-missing values on the independent and dependent variables.

Table 1 gives a summary of the quantile values of the variables involved in the regression, including the survey weights. The total expenditure has a maximum of 519,863.3, which is almost 30,000 times the minimum, 16.6. Although not as extreme as the total expenditure, the number of beds and the number of additions also have significant differences between their maxima and minima. Because the sample was selected using a PPS design, the sample weights were associated with the sizes of the mental health organizations, with a range from 1 to 158.86. The weights we use in analysis include a nonresponse adjustment which was done separately by design stratum. In some cases, units that were selected with certainty in the initial sample did not respond and some of the responding certainties had their weights adjusted to be larger than 1. A total of 157 organizations had a weight of 1 after the nonresponse adjustment.

Table 1
Quantiles of variables in SMHO regression

Variables	Quantiles				
	0%	25%	50%	75%	100%
Expenditure (1,000's)	16.6	2,932.5	6,240.5	11,842.6	519,863.3
# of Beds	0	6.5	36	93	2,405
# of Additions	0	558.5	1,410	2,406	79,808
Weights	1	1.42	2.48	7.76	158.86

In the regressions that follow, we have included the units with weights of 1 in standard error estimation rather than excluding them, as would be the approach for handling certainties in purely design-based estimation. Including the certainties is consistent with the idea that a superpopulation

model is being estimated and that slope coefficients would still have a variance even if a census were done. A sketch of the mathematical justification for doing this is model-dependent (not design-based) and is given in the Appendix.

Figure 1 shows scatterplots of expenditures versus beds and additions for the sample of 875 facilities (omitting one extremely large facility described below). In the first row, points are highlighted whose OLS leverage is greater than $2p/n = 0.007$. The second row shows bubbleplots with the relative size of the bubbles proportional to the weight of each case. High SW leverage points are highlighted using the same cutoff of 0.007. The distributions of the predictors are quite skewed as noted in Table 1. There is also one very large facility that is not shown in Figure 1 because it distorts the scale of the plot. That facility (denoted as observation 818 here) has (expenditures in 1,000's; beds; additions) = (\$519,863.3; 2,405; 79,808) and has a survey weight of 2.22. (Observation 818 was one of the cases noted earlier that was a certainty in the initial sample but received a nonresponse adjustment, and, thus, had a final weight larger than 1.) Because its data values are far out of line with those of the other organizations, this point has the potential to affect estimates.

Table 2 reports the twenty observations with the largest SW leverages. The values of the leverages range from 0.022 to 0.389, substantially greater than the level of the rough rule of thumb 0.007. This table also shows, for these twenty cases, the OLS unweighted leverages, the ratio of individual sample weight to average sample weight and the relative absolute distance between individual X values and their weighted means. We note that unit 818 has the highest weighted and unweighted leverages, mainly resulting from its extremely large number of beds and number of additions. Since this case has a less-than-average sample weight, the OLS leverage is even larger than the weighted one. There are other similar cases such as units 271, 179, 820, 157, 163, 156, and 154, which are associated with either extreme number of beds, or extreme number of additions, or both – but have small weights. Another type of outlier results from extreme sample weights, even if the values of their auxiliary variables are not very distinct from others. Units 672, 613, 711, 801, and 611 all have sample weights more than 15 times the average weight. Their weighted leverages are identified as large, whereas the unweighted leverages are not. There is also a noticeable gap between the weighted leverages for case 331 ($h_{ii} = 0.075$) and for case 271 ($h_{ii} = 0.046$).

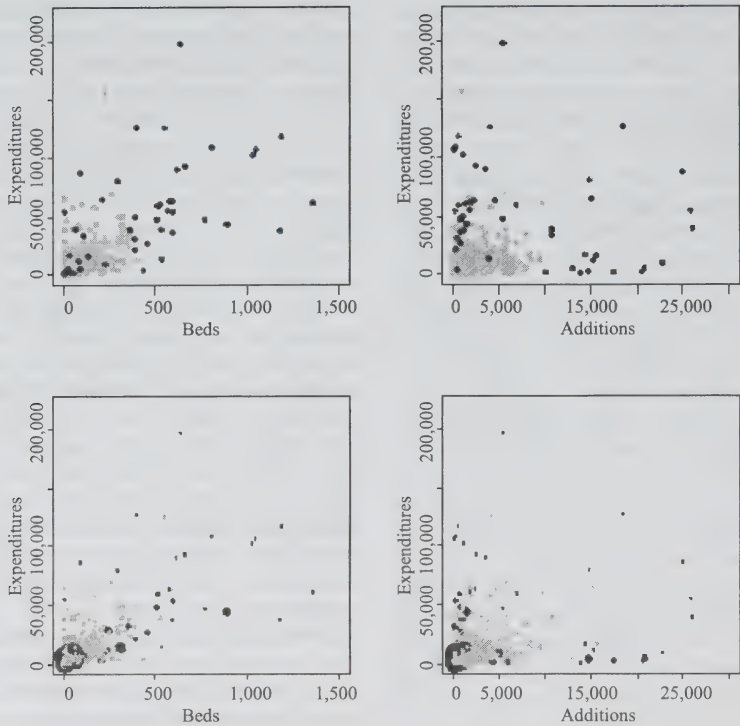


Figure 1 Scatterplots of expenditures versus beds and additions. High leverage points based on OLS (SW) are highlighted in top (bottom) row

Table 2
Observations with 20 largest survey weighted leverages

Obs ID	OLS h_{ii}	Weighted h_{ii}	Weights	Beds	Additions
			w_i / \bar{w}	$ x_{1i} - \bar{x}_1^w / \bar{x}_1^w$	$ x_{2i} - \bar{x}_2^w / \bar{x}_2^w$
818	0.513	0.389	0.3	49.3	64.7
189	0.037	0.245	3.4	17.7	0.3
346	0.035	0.157	2.2	0.6	16.1
366	0.017	0.105	3.0	0.7	11.1
331	0.024	0.075	1.5	0.1	13.4
271	0.068	0.046	0.4	23.7	0.0
830	0.004	0.045	5.8	5.4	0.1
628	0.056	0.045	0.4	1.0	20.3
179	0.089	0.038	0.2	27.4	0.5
672	0.002	0.034	24.2	1.0	0.8
820	0.048	0.034	0.3	0.8	19.6
207	0.012	0.030	1.3	9.5	0.3
157	0.069	0.030	0.2	23.8	0.5
163	0.017	0.027	0.8	11.4	0.8
613	0.002	0.026	18.5	1.0	0.7
711	0.002	0.024	16.8	1.0	0.9
801	0.002	0.024	17.5	0.6	0.9
156	0.055	0.023	0.2	20.9	0.9
611	0.002	0.023	15.9	1.0	0.8
154	0.051	0.022	0.2	20.5	0.1
			$\bar{w} = 6.57$	$\bar{x}_1^w = 47.83$	$\bar{x}_2^w = 1,214.13$

Note: observation ID is the line number of an observation in the sample.

Sizes of the sample weights can make analysts reach different conclusions when they use weighted or unweighted leverages to identify potentially influential observations. Figure 2 shows a scatterplot of weighted leverages versus unweighted ones. The two reference lines were drawn at values of 0.007. Observation 818 is omitted since it would again distort the scale of the graph. Clearly, the high leverage points identified by the SW method only, located in area A, have significantly larger weights than the points in area B, which are identified by the OLS method only.

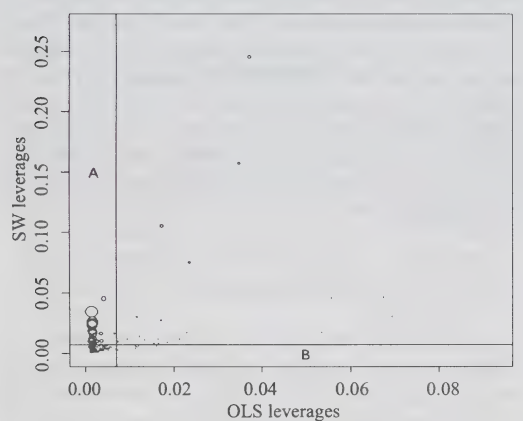


Figure 2 Plot of survey weighted leverages versus OLS unweighted leverages

Given that some potentially influential cases have been identified, the next step is to see what effect they have on parameter estimates. Table 3 shows the OLS and SW parameter estimates using all cases. Table 4 lists the OLS and SW estimates (i) omitting high leverage cases and (ii) omitting observation 818. High leverage points are those with $h_{ii} > 0.007$. However, note that different sets of points are high leverage in OLS and SW regressions. The standard errors are estimated via the usual OLS formula and the sandwich estimator (Binder 1983) for the SW estimates.

Comparing Tables 3 and 4, we see that the OLS estimates change substantially after high leverage points are deleted (section (i) of Table 4). The OLS intercept, which is significant in both tables, jumps from negative to positive. The OLS slope for beds drops by about 26% (94.16 to 69.27) when the high leverage points are dropped. The decrease is about 59% for the slope for additions. The SW estimates for beds and additions are also sensitive to the high leverage points with the slopes decreasing by 7% and 46% respectively. In all cases, the slopes are significant so

that the qualitative conclusion that expenditures is related to beds and additions holds with or without the high leverage points. However, predicted values will be quite different before and after omitting these points.

The standard errors (SE's) also decrease substantially when the high leverage points are omitted. For example, the SW standard error for beds drops from 13.14 to 6.75 (a 49% reduction); the SE for additions drops from 0.76 to 0.21 (a 72% reduction). This is due to some points with extreme weights being removed in the SW regression. In contrast, the SE's for the OLS estimates actually increase when the OLS high leverage points are omitted because the sample variance of the x 's decreases. This is another illustration of the considerable differences that can occur when applying the same type of diagnostic to OLS and SW regressions.

Table 3
OLS and SW parameter estimates of SMHO regression using all 875 sample cases

Independent Variables	OLS Estimation			SW Estimation		
	Coefficient	SE	t	Coefficient	SE	t
Intercept	-1,201.73	526.19	-2.28	514.08	1,157.71	0.44
# of Beds	94.16	3.03	31.08	81.23	13.14	6.18
# of Additions	2.31	0.13	18.50	1.84	0.76	2.43

Table 4
OLS and SW parameter estimates after from SMHO regression

Independent Variables	OLS Estimation			SW Estimation		
	Coefficient	SE	t	Coefficient	SE	t
(i) Deleting observations with leverages greater than 0.007						
Intercept	2,987.55	490.54	6.09	1,993.86	353.71	5.64
# of Beds	69.27	4.35	15.94	75.82	6.75	11.23
# of Additions	0.95	0.20	4.71	1.00	0.21	4.73
(ii) Deleting observation 818						
Intercept	1,979.51	537.93	3.68	2,281.17	460.35	4.96
# of Beds	81.80	2.92	27.98	68.69	8.04	8.54
# of Additions	1.19	0.14	8.41	0.79	0.29	2.75

Because point 818 is so obviously extreme, we also fitted the regression after dropping only that observation. The results are shown in section (ii) of Table 4. Omitting that single point causes noticeable changes in both OLS and SW parameter estimates. This also illustrates that a single point can affect the standard errors for estimated slopes in a survey-weighted regression, as is also the case in OLS. Observation 818 has a large residual (see Figure 3); omitting it results in the SE for Beds dropping from 13.14 in Table 3 to 8.04 in Table 4. Note that if unit 818 had a large weight, then its residual would likely be smaller since it would have more affect on the fit. If so, the SE could actually be smaller when unit 818 is included.

Another point to be gleaned from Tables 3 and 4 is that the OLS and SW estimates are much closer to each other after the high leverage points are dropped than they are before. As shown in Table 5, the OLS estimates are 16 and 26% larger than the SW estimates with all points but are 9 and 5% less than SW after dropping points.

Table 5
Ratios of OLS and SW parameter estimates before and after deleting observations with leverages greater than 0.007 from SMHO regression

	Ratio of OLS to SW estimates	
	With all points	Dropping high leverage points
Beds	1.16	0.91
Additions	1.26	0.95

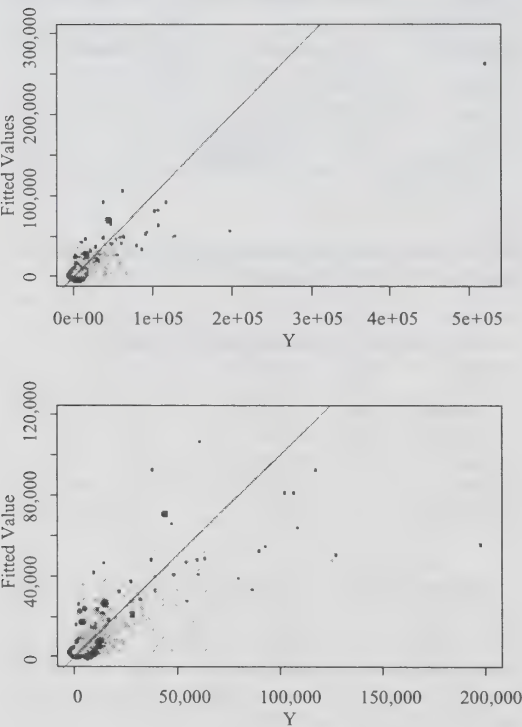


Figure 3 Plot of fitted values versus Y values. Reference line is drawn at $Y = \hat{Y}$. The upper panel includes all points. The lower panel omits the extreme observation 818. High leverage points based on SW are solid, dark circles in each panel

Leverages are usually combined with residuals to determine which points are influential in fitting the regression model because residuals can be used to detect discrepant Y values. A scatterplot of fitted values from the SW regression versus the Y values is shown in Figure 3. The high leverage points are labeled as dark solid circles. The vertical distances from the points to the 45 degree line imply the sizes of the residuals. The upper panel includes all 875 sample points; the lower panel omits observation 818 to provide better resolution for the remaining points. Note that some observations have high leverages and small residuals, while others have low leverages and large residuals. The influence of these points on the regression can be further investigated using various tools that we will not cover here. For example, Cook’s distance, implicitly involving the leverage and residual, is designed to measure the effect of deleting a single observation on the overall parameter estimates. The adaptation of some basic OLS diagnostic statistics to survey data, such as DFBETAS and DFFITS, has been discussed under a single stage sampling design in Li and Valliant (2006).

6. Conclusion

Leverages and residuals are essential components of diagnostic statistics intended to identify substantial influence of a single observation or a group of observations on a fitted linear model. Survey data sets can contain influential observations whether one argues that the sample design is ignorable and ordinary least squares can be used, or that the design must be accounted for and survey weights used. The points that are influential in the two cases are not necessarily the same, as illustrated here.

Once high leverage points are identified, an important question is how to deal with them for inference. Two options are to down-weight them or drop them from model-fitting entirely. Down-weighting seems unsatisfactory in general since a point can have a high leverage not because of a large weight but rather due to having one or more unusual X’s. Down-weighting may be sensible from a model-based point-of-view, assuming the model itself is correctly specified. However, the design-based idea of estimating a census parameter may then be lost. If a point has a large leverage because of extreme X’s, then it may not follow the model at all and should be dropped.

However, using a mechanical procedure that automatically drops many influential observations with high leverages can lead to standard error estimates that are too small, resulting in confidence intervals that cover at less than the nominal rates and in inflated Type I error rates in hypothesis tests (Li 2007). This phenomenon is similar to well-known problems in stepwise regression (Hurvich and

Tsai 1990, Zhang 1992). Thus, a useful research topic appears to be developing inferential procedures for constructing confidence intervals and conducting hypothesis tests that account for the effects of dropping or down-weighting points.

For complex survey data, the hat matrix involves no design features except for sample weights and can be used to identify cases that have atypical weights or predictor values. Other diagnostic statistics, like Cook's D, do contain variance estimates that need to account for complex sample design features such as stratification and clustering. The adaptation and extension of additional diagnostic approaches for survey analysis will be explored in the future.

7. Acknowledgement

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 0617081. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank the editor and referees for their thoughtful comments, which improved the paper considerably.

Appendix

Inclusion of certainties in standard error estimation

In the empirical study in section 5, we included certainty units in the standard error calculations. The justification for doing this is sketched here. Under the general model (3), the model variance of $\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$, the estimator used in the empirical study, is $\text{var}_M(\hat{\beta}) = \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{X} \mathbf{A}^{-1} \sigma^2$ where $\mathbf{A} = \mathbf{X}^T \mathbf{W} \mathbf{X}$ and $\mathbf{V} = \text{diag}(v_i)_{i \in s}$. The sandwich variance estimator used in the study reported in section 5 is defined as

$$v(\hat{\beta}) = \mathbf{A}^{-1} \frac{n}{n-1} \sum_{i \in s} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T \mathbf{A}^{-1} \quad (4)$$

where $\mathbf{z}_i = w_i \mathbf{e}_i \mathbf{x}_i$ with $\mathbf{e}_i = Y_i - \mathbf{x}_i^T \hat{\beta}$ and $\bar{\mathbf{z}} = \sum_{i \in s} w_i \mathbf{e}_i \mathbf{x}_i / n$. This estimator is design consistent (see Binder 1983) in single-stage sampling if units are sampled with replacement with probabilities equal to w_i^{-1} , and there are no certainty units. If the sample contains certainties, the formula for $v(\hat{\beta})$ would be modified to estimate the design-based variance: certainties would be excluded from the sums in (4) and $\bar{\mathbf{z}}$, and n would be changed to n_{nc} , the number of non-certainties. In the extreme case of a census, the design-based variance estimator would reduce to zero.

The estimator in (4) is approximately model-unbiased under (3) regardless of whether the sample contains certainties or not. The middle matrix in (4) can be expanded as $\sum_{i \in s} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T = \sum_{i \in s} \mathbf{z}_i \mathbf{z}_i^T - n \bar{\mathbf{z}} \bar{\mathbf{z}}^T$. Assuming that $e_i \approx Y_i - \mathbf{x}_i^T \hat{\beta}$, the model expectation under (3) of the first term is $E_M(\sum_{i \in s} \mathbf{z}_i \mathbf{z}_i^T) = \mathbf{X}^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{X} \sigma^2$ while $E_M(n \bar{\mathbf{z}} \bar{\mathbf{z}}^T) = n^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{X} \sigma^2$. Substituting these expectations gives $E_M[v(\hat{\beta})] = \text{var}_M(\hat{\beta})$, which holds even when some units are certainties. This also shows that $v(\hat{\beta})$ is robust in the sense of properly reflecting the contribution of heteroscedastic variances in (3) to the model-variance of $\hat{\beta}$ even though \mathbf{V} may be unknown and not accounted for in the estimation of $\hat{\beta}$.

References

- Beaumont, J.-F., and Alavi, A. (2004). Robust Generalized Regression Estimation. *Survey Methodology*, 30, 195-208.
- Belsley, D.A., Kuh, E. and Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Inc.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters, Chapter 3 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.
- Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- Chambers, R.L., Dorfman, A.H. and Sverchkov, M.Y. (2003). Nonparametric regression with complex survey data, Chapter 11 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.
- Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.
- Choudhry, G. (2000). The 1998 Survey of Mental Health Organizations Survey Design. Westat technical report prepared for Center for Mental Health Services, Substance Abuse and Mental Health Services Administration (SAMHSA), available by request to SAMHSA.
- Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.
- Gwet, J., and Rivest, L. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

- Hoaglin, D.C., and Welsch, R.E. (1978). The hat matrix in regression and ANOVA (Corr: 78V32 p146). *The American Statistician*, 32, 17-22.
- Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.
- Hurvich, C.M., and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.
- Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.
- Li, J. (2007). *Regression Diagnostics for Complex Survey Data: Identification of Influential Observations*. Unpublished doctoral dissertation, University of Maryland.
- Li, J., and Valliant, R. (2006). Influence analysis in linear regression with sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3330-3337.
- Manderscheid, R.W., and Henderson, M.J. (2002). Mental Health, United States, 2002. DHHS Publication No. SMA04-3938. Rockville MD USA: Substance Abuse and Mental Health Services Administration. available at <http://mentalhealth.samhsa.gov/publications/allpubs/SMA04-3938/AppendixA.asp>
- Moreno-Rebollo, J.L., Muñoz-Reyes, A. and Muñoz-Pichardo, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.
- Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.
- Potter, F.J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.
- Potter, F.J. (1993). The effect of weight trimming on nonlinear survey estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.
- Skinner, C.J. (2003). Introduction to Part B, Chapter 6 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.
- Smith, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14, 143-152.
- Smith, T.M.F. (1989). Introduction to Part B, Chapter 6 in *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons, Inc.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- Weisberg, S. (2005). *Applied Linear Regression*, Third Edition. New York: John Wiley & Sons, Inc.
- Welsh, A.H., and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B, Methodological*, 60, 413-428.
- Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting influential clusters in surveys: Application to the 1990 post enumeration Survey. *Journal of the American Statistical Association*, 96, 858-869.
- Zhang, P. (1992). Influence after variable selection in linear regression models. *Biometrika*, 79, 741-746.

A practical bootstrap method for testing hypotheses from survey data

Jean-François Beaumont and Cynthia Bocci¹

Abstract

The bootstrap technique is becoming more and more popular in sample surveys conducted by national statistical agencies. In most of its implementations, several sets of bootstrap weights accompany the survey microdata file given to analysts. So far, the use of the technique in practice seems to have been mostly limited to variance estimation problems. In this paper, we propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The probability sampling design used to select the sample may be informative or not. Our method uses model-based test statistics that incorporate the survey weights. Such statistics are usually easily obtained using classical software packages. We approximate the distribution under the null hypothesis of these weighted model-based statistics by using bootstrap weights. An advantage of our bootstrap method over existing methods of hypothesis testing with survey data is that, once sets of bootstrap weights are provided to analysts, it is very easy to apply even when no specialized software dealing with complex surveys is available. Also, our simulation results suggest that, overall, it performs similarly to the Rao-Scott procedure and better than the Wald and Bonferroni procedures when testing hypotheses about a vector of linear regression model parameters.

Key Words: Bootstrap weights; Analysis of survey data; Hypothesis testing; Informative sampling; Linear regression; Model parameters.

1. Introduction

The bootstrap technique is becoming more and more popular in sample surveys conducted by national statistical agencies. The main reasons seem to be that it can easily deal with several situations that would be difficult to handle otherwise (*e.g.*, nonresponse weight adjustment, calibration, non-smooth statistics, *etc.*) and that it is convenient for analysts. In most of its implementations, several sets of bootstrap weights accompany the survey microdata file given to analysts; no other design information is provided. These weights are usually obtained by assuming that the first-stage sampling fractions are small enough that a without-replacement sampling design can be accurately approximated by a with-replacement sampling design. The reader is referred to Rao, Wu and Yue (1992) for a succinct but clear description of a method to construct bootstrap weights under this assumption when a stratified multistage sampling design has been used.

So far, the use of the technique in practice seems to have been mostly limited to variance estimation problems (*e.g.*, Langlet, Faucher and Lesage 2003; Yeo, Mantel, and Liu 1999; and Hughes and Brodsky 1994). On the research side, efforts have been mainly oriented towards finding an appropriate bootstrap methodology for variance estimation when the sample is drawn without replacement from a finite population (see Sitter 1992; or Shao and Tu 1995, Chapter 6, for a review of methods). Some authors have also studied the problem of determining bootstrap confidence intervals

for a finite population parameter (*e.g.*, Rao and Wu 1988; Kovar, Rao and Wu 1988; Sitter 1992; and Rao *et al.* 1992). To our knowledge, there does not seem to be any literature on hypothesis testing using the bootstrap technique in survey sampling although this problem has been studied in the context of classical statistics. The reader is referred to Hall and Wilson (1991) for a discussion on bootstrap tests of hypotheses and to Efron and Tibshirani (1993) for an excellent account of the bootstrap technique in classical statistics. It is worth noting the work of Graubard, Korn and Midthune (1997) who applied the classical parametric bootstrap method to survey data in order to test the fit of a logistic regression model. Their procedure is valid when sampling is not informative.

The problem of hypothesis testing from complex survey data has been well studied in the last 30 years (*e.g.*, Rao and Scott 1981; Fay 1985; Thomas and Rao 1987; Korn and Graubard 1990; Korn and Graubard 1991; Graubard and Korn 1993; Thomas, Singh and Roberts 1996; and Rao and Thomas 2003). However, except perhaps for estimating unknown variances/covariances involved in these methods, the bootstrap technique has apparently not yet been considered for testing hypotheses. The goal of this paper is thus to propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The probability sampling design used to select the sample may be informative or not. Informally speaking, sampling is informative when the model that holds for the selected

1. Jean-François Beaumont, Statistics Canada, Statistical Research and Innovation Division, Tunney's Pasture, R.H. Coats building, 16th floor, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Jean-Francois.Beaumont@statcan.gc.ca; Cynthia Bocci, Statistics Canada, Business Survey Methods Division, Tunney's Pasture, R.H. Coats building, 11th floor, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Cynthia.Bocci@statcan.gc.ca.

sample is different from the model that holds for the whole population; otherwise sampling is not informative.

Our method uses model-based test statistics that incorporate the survey weights. Such statistics are usually easily obtained using classical software packages. We approximate the distribution under the null hypothesis of these weighted model-based statistics by using bootstrap weights. An advantage of our bootstrap method over existing methods of hypothesis testing with survey data is that, once sets of bootstrap weights are provided to analysts, it is very easy to apply even when no specialized software dealing with complex surveys is available.

We introduce notation and the problem in section 2. In section 3, we describe and justify our proposed bootstrap methodology for testing hypotheses with survey data. A linear regression example is given in section 4 to illustrate the theory. We briefly describe the alternative Rao-Scott (Rao and Scott 1981), Wald and Bonferroni procedures in section 5 when testing hypotheses about a vector of linear regression model parameters. They are evaluated in section 6 and compared to our proposed bootstrap procedure through a simulation study. Finally, we conclude in the last section with a short summary and discussion.

2. Preliminaries

We assume that a finite population U of size N has been generated according to a model, specified by the analyst, that describes the conditional distribution $F(y_U | X_U; \beta, \theta)$. The N -vector y_U contains the population values of a dependent variable y , X_U is an N -row matrix that contains the population values of a vector of independent variables x , β is an r -vector of unknown model parameters and θ is a potential vector of additional unknown model parameters. We are interested in testing hypotheses about β but not θ . We also assume that, if the entire population U could be observed, a test statistic $t(U; c)$ would be used to test the multiple linear hypothesis $H_0: H\beta = c$ against the alternative hypothesis $H_1: H\beta \neq c$. The $Q \times r$ matrix H is used to define the hypothesis to be tested and c is a Q -vector of constants specified by the analyst. Ideally, $t(U; c)$ is asymptotically pivotal; i.e., it has an asymptotic distribution that does not depend on any unknown parameter. We consider statistics that have the following quadratic form:

$$t(U; c) = (H\hat{\beta}_U - c)' \{A(U)\}^{-1} (H\hat{\beta}_U - c), \quad (2.1)$$

where $\hat{\beta}_U$ is a consistent estimator of β under the model and $A(U)$ is some scaling matrix. Typically, $A(U)$ is symmetric and positive definite.

As an illustrative example, let us assume that y_k , for all population units $k \in U$, are independently and identically

distributed random variables with mean β and variance θ and that we are interested in testing the null hypothesis $H_0: \beta = c$. In this example, $Q = 1$, $r = 1$, $H = 1$ and $X_U = 1_U$, where 1_U is a population vector of one's. A common test statistic for this problem is

$$t(U; c) = \frac{(\hat{\beta}_U - c)^2}{\hat{\theta}_U / N}, \quad (2.2)$$

where $\hat{\beta}_U = \sum_{k \in U} y_k / N$ and $\hat{\theta}_U = \sum_{k \in U} (y_k - \hat{\beta}_U)^2 / (N - 1)$. The statistic (2.2) has the same form as (2.1) if we let $A(U) = \hat{\theta}_U / N$. This statistic is usually assumed to follow the distribution χ^2_1 or $F_{1, N-1}$ under the null hypothesis.

As is typically the case, a random sample s of size n is selected from the finite population U according to a given probability sampling design $p(s)$. Since the dependent variable y and, possibly, the independent variables x are not observed for nonsample units, we may want to use the statistic $t(s; c)$ instead of $t(U; c)$. In the above example, this would lead to $t(s; c) = n(\hat{\beta}_s - c)^2 / \hat{\theta}_s$, where $\hat{\beta}_s = \sum_{k \in s} y_k / n$ and $\hat{\theta}_s = \sum_{k \in s} (y_k - \hat{\beta}_s)^2 / (n - 1)$. However, if sampling is informative with respect to the model, it may be more appropriate and is undoubtedly more common to use a weighted test statistic of the form

$$\hat{t}(s, w_s; c) = (H\hat{\beta}_{ws} - c)' \{\hat{A}(s, w_s)\}^{-1} (H\hat{\beta}_{ws} - c). \quad (2.3)$$

The n -vector w_s contains the survey weight of sample unit k in its k^{th} element, denoted by w_k , $\hat{\beta}_{ws}$ is a weighted estimator for β and $\hat{A}(s, w_s)$ is a weighted analogue to $A(s)$ in that each sample unit k is weighted by its survey weight w_k whereas there is no weighting with $A(s)$. We thus have $\hat{A}(s, 1_s) = A(s)$, where 1_s is a sample vector of one's. As a result, the statistic $\hat{t}(s, w_s; c)$ is also a weighted analogue to $t(s; c)$ and we have $\hat{t}(s, 1_s; c) = t(s; c)$. If the statistic $t(s; c)$ can be computed using some classical software package, not necessarily developed to handle survey data, the statistic $\hat{t}(s, w_s; c)$ can also be computed using the same software package provided that it can allow each observation to be weighted by its survey weight.

Typically, the survey weight w_k , for a unit $k \in s$, is equal to the inverse of its selection probability, which may then be calibrated to account for known external information (e.g., Deville and Särndal 1992). We assume that the sampling design and the survey weights are constructed so that the following two assumptions hold:

Assumption 1: $\sqrt{n}(\hat{\beta}_{ws} - \beta) \xrightarrow{mp} N(0, \Sigma)$, where \xrightarrow{mp} denotes convergence in distribution under the model and the sampling design, and Σ is the asymptotic variance-covariance matrix of $\sqrt{n}\hat{\beta}_{ws}$ under the model and the sampling design. The notation “ m ” stands for the model while the notation “ p ” stands for the probability sampling design.

Assumption 2: $n\hat{\mathbf{A}}(s, \mathbf{w}_s)$ is symmetric, positive definite and mp -consistent for some fixed symmetric positive definite scaling matrix $\hat{\mathbf{A}}$.

Note that assumption 2 does not require $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ to be p -consistent for $\mathbf{A}(U)$. Indeed, $NA(U)$ will be typically m -consistent for $\hat{\mathbf{A}}$. Other choices could replace the weighted scaling matrix $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ in (2.3). For instance, it could be replaced by an estimator of the design variance of $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ under simple random sampling (e.g., Rao and Scott 1981). An alternative choice is the common Wald statistic. It is obtained by replacing $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ in (2.3) by $\hat{\mathbf{V}}_{mp}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, which is an mp -consistent estimator of $\mathbf{V}_{mp}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$; the variance of $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ evaluated with respect to the model and the sampling design. As pointed out in the paragraph below (2.3), an advantage of using a scaling matrix $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ such that $\hat{\mathbf{A}}(s, \mathbf{1}_s) = \mathbf{A}(s)$ is that the resulting test statistic $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ can then be directly computed using classical software packages provided that they allow each observation to be weighted by its survey weight. It is thus more convenient for the users of survey data.

Continuing the above example, we may define our weighted test statistic as

$$\hat{t}(s, \mathbf{w}_s; \mathbf{c}) = \frac{(\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})^2}{\{(\hat{N} - 1)/(n - 1)\} (\hat{\boldsymbol{\theta}}_{ws}/\hat{N})}, \quad (2.4)$$

where $\hat{N} = \sum_{k \in s} w_k$, $\hat{\boldsymbol{\beta}}_{ws} = \sum_{k \in s} w_k y_k / \sum_{k \in s} w_k$ and $\hat{\boldsymbol{\theta}}_{ws} = \sum_{k \in s} w_k (y_k - \hat{\boldsymbol{\beta}}_{ws})^2 / (\hat{N} - 1)$. In (2.4), the underlying weighted scaling matrix is $\hat{\mathbf{A}}(s, \mathbf{w}_s) = \{(\hat{N} - 1)/(n - 1)\} (\hat{\boldsymbol{\theta}}_{ws}/\hat{N})$, which does not depend on the way the weights are scaled. If they are rescaled so that $\sum_{k \in s} w_k = n$, which is typically done by analysts, then the factor $(\hat{N} - 1)/(n - 1)$ vanishes. The role of this factor, along with other regularity conditions, is to satisfy assumption 2. If the SAS® System is chosen, the test statistic (2.4) is obtained by using the WEIGHT statement in standard procedures. When the null hypothesis is true, it is well known that (2.4) unfortunately does not follow the distribution χ^2_1 or $F_{1, n-1}$ under the model and the sampling design.

To obtain a valid test procedure, we need to approximate the distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ under the null hypothesis. This can be achieved by using the following result:

Result 1: $\hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^Q \lambda_q \Omega_q$, where λ_q , for $q = 1, \dots, Q$, are the eigenvalues of $\boldsymbol{\Lambda} = (\hat{\mathbf{A}}^{-1})(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$ and Ω_q are independent chi-square random variables with one degree of freedom.

The proof of result 1 uses assumptions 1 and 2 and is given in the appendix. When the null hypothesis is true (i.e., $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$), we thus have

$$\hat{t}(s, \mathbf{w}_s; \mathbf{c}) \xrightarrow{mp} \sum_{q=1}^Q \lambda_q \Omega_q. \quad (2.5)$$

Rao and Scott (1981) used a similar result to construct their test procedures. They approximated a distribution like (2.5) by a scaled chi-square distribution that matches the estimated first two moments of the right-hand side of (2.5). Instead, we approximate the distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ under the null hypothesis by using bootstrap weights. This is described in the next section.

Before giving details of our test procedure, it is useful to note that $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ in (2.3) can be written as

$$\begin{aligned} \hat{t}(s, \mathbf{w}_s; \mathbf{c}) &= \hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta}) \\ &+ 2(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})' \{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1} (\mathbf{H}\boldsymbol{\beta} - \mathbf{c}) \\ &+ (\mathbf{H}\boldsymbol{\beta} - \mathbf{c})' \{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1} (\mathbf{H}\boldsymbol{\beta} - \mathbf{c}). \end{aligned} \quad (2.6)$$

Under the null hypothesis, the last two terms on the right-hand side of (2.6) vanish and we have $\hat{t}(s, \mathbf{w}_s; \mathbf{c}) = \hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta})$. When the null hypothesis is false, the third term on the right-hand side of (2.6) dominates the others as the sample size increases since the first, second and third terms are $O_p(1)$, $O_p(\sqrt{n})$ and $O_p(n)$ respectively, provided that assumptions 1 and 2 hold. Also, since $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ is positive definite, the third term is always positive. Therefore, a large positive observed value of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ compared to a large percentile of the distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta})$ is an indication that the null hypothesis may be wrong.

3. The proposed bootstrap method

Let w_k^* denote a random bootstrap weight for unit k , obtained using some bootstrap procedure such as that of Rao *et al.* (1992), and let \mathbf{w}_s^* be the n -vector that contains the random bootstrap weight w_k^* in its k^{th} element. The bootstrap estimator $\hat{\boldsymbol{\beta}}_{ws}^*$ is obtained similarly to $\hat{\boldsymbol{\beta}}_{ws}$ by replacing the survey weight w_k by its bootstrap version w_k^* for each sample unit. We also denote by \mathbf{w}_s^{*b} , for $b = 1, \dots, B$, the B n -vectors containing the bootstrap weights w_k^{*b} in their k^{th} element. These B vectors are drawn independently and have the same distribution as \mathbf{w}_s^* ; this distribution is called the bootstrap distribution and is denoted by the symbol “*”. The b^{th} bootstrap estimator $\hat{\boldsymbol{\beta}}_{ws}^{*b}$ is defined in an obvious manner.

Before describing our bootstrap test procedure, we first introduce three additional assumptions related to the construction of the bootstrap weights:

Assumption 3: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{ws}^* - \hat{\boldsymbol{\beta}}_{ws}) \xrightarrow{*} N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, where $\xrightarrow{*}$ denotes convergence in bootstrap distribution and

$\hat{\Sigma}$ is the asymptotic bootstrap variance-covariance matrix of $\sqrt{n} \hat{\beta}_{ws}$.

Assumption 4: $n\hat{\mathbf{A}}(s, \mathbf{w}_s^*)$ is $*$ -consistent for $n\hat{\mathbf{A}}(s, \mathbf{w}_s)$.

Assumption 5: $\hat{\Sigma}$ is mp -consistent for Σ .

Assumptions 3 and 4 are bootstrap analogues to assumptions 1 and 2 and should be satisfied with most bootstrap methods (e.g., those described in the review paper by Sitter 1992) and models (e.g., linear regression model, logistic regression model, etc.). The reader is referred to Shao and Tu (1995, Chapter 6; in particular section 6.4.4) for greater detail.

A comment is in order about assumption 5. This assumption is equivalent to requiring that the bootstrap variance $\mathbf{V}_m(\hat{\beta}_{ws}^*)$ be mp -consistent for

$$\mathbf{V}_{mp}(\hat{\beta}_{ws}) = \mathbf{E}_m \mathbf{V}_p(\hat{\beta}_{ws}) + \mathbf{V}_m \mathbf{E}_p(\hat{\beta}_{ws}). \quad (3.1)$$

This means that the bootstrap distribution must reflect the variability due to both the model and the sampling design. Unfortunately, standard design-based bootstrap methods reflect only the variability due to the sampling design so that they only track the first term of the right-hand side of (3.1). Thus, these bootstrap methods do not satisfy assumption 5 in general. However, when the overall sampling fraction n/N is negligible, the second term of the right-hand side of (3.1) becomes negligible (e.g., see Binder and Roberts 2003) so that the approximation $\mathbf{V}_{mp}(\hat{\beta}_{ws}) \approx \mathbf{E}_m \mathbf{V}_p(\hat{\beta}_{ws})$ is appropriate and design-based bootstrap methods can be used. In many household surveys, the overall sampling fraction is actually quite small. Indeed, bootstrap weights are often obtained under the assumption that the first-stage sampling fractions are small (e.g., Rao *et al.* 1992). Developing bootstrap procedures that capture both terms of (3.1) is an area for future research.

Under assumptions 3 and 4, we obtain our second result:

Result 2: $\hat{t}(s, \mathbf{w}_s^*; \mathbf{H}\hat{\beta}_{ws}) \xrightarrow{*} \sum_{q=1}^Q \hat{\lambda}_q \Omega_q$, where $\hat{\lambda}_q$, for $q = 1, \dots, Q$, are the eigenvalues of $\hat{\mathbf{A}} = [n\hat{\mathbf{A}}(s, \mathbf{w}_s)]^{-1}(\mathbf{H}\hat{\Sigma}\mathbf{H}')$ and Ω_q are again independent chi-square random variables with one degree of freedom.

The proof of result 2 is omitted as it is very similar to the proof of result 1 given in the appendix. From assumptions 2 and 5, $\hat{\mathbf{A}}$ is mp -consistent for \mathbf{A} . Thus, using results 1 and 2, the bootstrap distribution of $\hat{t}(s, \mathbf{w}_s^*; \mathbf{H}\hat{\beta}_{ws})$ is asymptotically the same as the mp -distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{H}\hat{\beta})$, which is itself the same as the mp -distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ under the null hypothesis; the distribution that we want to approximate. This suggests the following bootstrap test procedure:

- i) Obtain bootstrap weights, w_k^{*b} , for $k \in s$ and $b = 1, \dots, B$.

- ii) Compute $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\beta}_{ws}^b)$, for $b = 1, \dots, B$.
- iii) Since a large value of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ leads to rejecting the null hypothesis, compute the observed significance level (p -value) as

$$\frac{\#\{\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\beta}_{ws}^b) > \hat{t}(s, \mathbf{w}_s; \mathbf{c})\}}{B}.$$

The null hypothesis is rejected if this value is lower than the significance level α (e.g., 5%).

Note that the statistic to be bootstrapped is $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\beta}_{ws}^b)$ and not $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{c})$. The use of the latter would not properly reflect the distribution under the null hypothesis and would thus violate the first guideline in Hall and Wilson (1991).

If $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ is pivotal then the second guideline of Hall and Wilson (1991) is also satisfied. The fact that $t(U; \mathbf{c})$ is asymptotically pivotal certainly helps in obtaining a better bootstrap test procedure. However, it does not unfortunately guarantee that $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ is also asymptotically pivotal, particularly when sampling is informative. Nevertheless, failure to use a pivotal statistic does not invalidate the above test procedure and may not reduce its power. But, it may reduce the level accuracy of the test. As pointed out by Hall and Wilson (1991), it is sometimes appropriate to disregard the second guideline. The main advantage of using the simple (possibly non-pivotal) statistic $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ in (2.3) and the bootstrap statistic $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\beta}_{ws}^b)$ is that, once bootstrap weights have been provided on the microdata file, these statistics are easily obtained using classical software packages that ignore sampling design features. Moreover, we show in section 5, through a simulation study, that our bootstrap test procedure performs similarly to the Rao-Scott procedure and better than the Wald and Bonferroni procedures.

4. A linear regression example

To better illustrate the theory in a practical context, let us now assume that, conditional on \mathbf{X}_U , the random variables y_k , for $k \in U$, are independently distributed with mean $\mathbf{E}_m(y_k | \mathbf{X}_U) = \mathbf{x}_k' \boldsymbol{\beta}$ and variance $\mathbf{V}_m(y_k | \mathbf{X}_U) = \theta$, where \mathbf{x}_k is an r -vector of linearly independent variables for unit k . Recall that we are interested in testing the null hypothesis $H_0: \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ against the alternative hypothesis $H_1: \mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$. If the entire population could be observed, the common statistic

$$t(U; \mathbf{c}) =$$

$$\frac{(\mathbf{H}\hat{\beta}_U - \mathbf{c})' \left(\mathbf{H} \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{H}' \right)^{-1} (\mathbf{H}\hat{\beta}_U - \mathbf{c})}{Q \hat{\theta}_U} \quad (4.1)$$

could be used, where

$$\hat{\beta}_U = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k$$

and

$$\hat{\theta}_U = \frac{\sum_{k \in U} (y_k - \mathbf{x}_k' \hat{\beta}_U)^2}{N - r}.$$

The statistic $t(U; \mathbf{c})$ in (4.1) follows the distribution $F_{Q, N-r}$ under the null hypothesis. It reduces to (2.2) when $Q = r = \mathbf{H} = \mathbf{x}_k = 1$ in (4.1).

A weighted sample version of (4.1), which can be written in the form of (2.3), is

$$\hat{t}(s, \mathbf{w}_s; \mathbf{c}) =$$

$$\frac{(\mathbf{H}\hat{\beta}_{ws} - \mathbf{c})' \left(\mathbf{H} \left(\sum_{k \in s} w_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{H}' \right)^{-1} (\mathbf{H}\hat{\beta}_{ws} - \mathbf{c})}{Q \hat{\theta}_{ws} \{(\hat{N} - r)/(n - r)\}}, \quad (4.2)$$

where

$$\hat{\beta}_{ws} = \left(\sum_{k \in s} w_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} w_k \mathbf{x}_k y_k \quad (4.3)$$

and

$$\hat{\theta}_{ws} = \frac{\sum_{k \in s} w_k (y_k - \mathbf{x}_k' \hat{\beta}_{ws})^2}{\hat{N} - r}. \quad (4.4)$$

For instance, the statistic $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ in (4.2) could be obtained by using the WEIGHT statement in the procedure REG of SAS as long as $w_k > 0$, for $k \in s$. Note that it satisfies assumption 2 and does not depend on the way the weights are scaled. Again, if the weights are rescaled so that $\sum_{k \in s} w_k = n$, the factor $(\hat{N} - r)/(n - r)$ in (4.2) vanishes. The test statistic (4.2) reduces to (2.4) when $Q = r = \mathbf{H} = \mathbf{x}_k = 1$ in (4.2), (4.3) and (4.4). The bootstrap statistic $\hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$ as well as $\hat{\beta}_{ws^b}$ and $\hat{\theta}_{ws^b}$ are obtained similarly to $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$, $\hat{\beta}_{ws}$ and $\hat{\theta}_{ws}$ in (4.2), (4.3) and (4.4) respectively, except that w_k is replaced by w_k^b and \mathbf{c} is replaced by $\mathbf{H}\hat{\beta}_{ws}$.

Remark 1: Note that w_k^b is likely to be 0 for some units $k \in s$ (see, for example, Rao *et al.* 1992). In some software packages such as SAS, the number of observations used in the analysis of the b^{th} bootstrap replicate, n^b , is equal to the number of units $k \in s$ for which $w_k^b > 0$. Such software packages may use $n^b - r$ instead of $n - r$ when computing the bootstrap statistic $\hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$. One must thus make sure that $n - r$ is used and, if not, that the bootstrap statistic computed from these packages is properly adjusted before applying the proposed bootstrap test procedure. One way of avoiding this problem is to add a very small positive value (e.g., 1×10^{-10}) to each bootstrap

weight w_k^b , for $k \in s$, so that no observation is excluded from the computation of $\hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$.

Remark 2: Let us define the bootstrap statistic $\hat{t}_e(s, \mathbf{w}_s^b; \mathbf{0})$ by replacing y_k by $e_k = y_k - \mathbf{x}_k' \hat{\beta}_{ws}$ in $\hat{t}(s, \mathbf{w}_s^b; \mathbf{0})$, for each $k \in s$. It is not difficult to show that $\hat{t}_e(s, \mathbf{w}_s^b; \mathbf{0}) = \hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$ so that our bootstrap procedure can be implemented using either $\hat{t}_e(s, \mathbf{w}_s^b; \mathbf{0})$ or $\hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$ when a linear regression model is used. The former may sometimes be more convenient with some software packages. This was the case in our simulation study since the use of $\hat{t}_e(s, \mathbf{w}_s^b; \mathbf{0})$ allowed us to get rid of manually typing the values of $\mathbf{H}\hat{\beta}_{ws}$ for each selected sample. An informal explanation for the equality $\hat{t}_e(s, \mathbf{w}_s^b; \mathbf{0}) = \hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$ can be obtained by treating $\hat{\beta}_{ws}$ as a fixed quantity, which is actually the case under the bootstrap distribution. The bootstrap statistic $\hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$ can thus be interpreted as a statistic aiming at testing the null hypothesis $H_0: \mathbf{H}\beta = \mathbf{H}\hat{\beta}_{ws}$ or, alternatively, $H_0: \mathbf{H}\gamma = \mathbf{0}$, where $\gamma = \beta - \hat{\beta}_{ws}$. Still assuming that $\hat{\beta}_{ws}$ is fixed, we can rewrite our linear model $E_m(y_k | \mathbf{X}_U) = \mathbf{x}_k' \beta$ as $E_m(e_k | \mathbf{X}_U) = \mathbf{x}_k' \gamma$. These observations seem to imply that using the bootstrap statistic $\hat{t}_e(s, \mathbf{w}_s^b; \mathbf{0})$ is equivalent to using $\hat{t}(s, \mathbf{w}_s^b; \mathbf{H}\hat{\beta}_{ws})$, which is indeed true.

Remark 3: We have already mentioned that the WEIGHT statement is necessary to obtain a weighted statistic if the proposed bootstrap test procedure is implemented using the procedure REG of SAS. Also, the TEST statement is necessary to request the desired statistics to be produced and the "ODS OUTPUT TESTANOVA =" statement to save these requested statistics in a SAS dataset specified by the user.

5. Some alternative procedures for linear regression

In this section, we briefly describe some test procedures in the context of linear regression exposed in section 4; namely, two naïve procedures that are sometimes used in practice as well as specific implementations of the Rao-Scott, Wald and Bonferroni procedures. They will all be evaluated in the simulation study in section 6.

The Bonferroni, Wald and Rao-Scott procedures, described in sections 5.2, 5.3 and 5.4 respectively, all need an mp -consistent estimator $\hat{\mathbf{V}}_{mp}(\hat{\beta}_{ws})$ of $\mathbf{V}_{mp}(\hat{\beta}_{ws})$. In the simulation study in section 6, we have used the bootstrap variance estimator

$$\hat{\mathbf{V}}_{mp}(\hat{\beta}_{ws}) = \frac{\sum_{b=1}^B (\hat{\beta}_{ws^b} - \hat{\beta}_{ws})(\hat{\beta}_{ws^b} - \hat{\beta}_{ws})'}{B}. \quad (5.1)$$

It is worth noting that the validity of assumption 5 is thus not only required for our proposed bootstrap method but also for the Bonferroni, Wald and Rao-Scott methods.

5.1 Two naïve procedures

The weighted version of the naïve procedure consists of using the statistic $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ in (4.2), which is compared to the upper tail of the distribution $F_{Q, n-r}$. The unweighted version uses the statistic $\hat{t}(s, \mathbf{1}_r; \mathbf{c})$, which is again compared to the upper tail of the distribution $F_{Q, n-r}$. Both procedures are not expected to work well under informative sampling but are still often used in practice, especially the weighted version. Note that if sampling is not informative, the unweighted version, that ignores the sampling design, leads to a simple, valid and reasonably powerful test.

5.2 The Bonferroni procedure

The Bonferroni procedure was studied by Korn and Graubard (1990). It is simple to use and was shown to work well in their empirical study. To describe this procedure, let \mathbf{H}'_q represent the q^{th} row of \mathbf{H} and c_q the q^{th} element of \mathbf{c} . Then, compute the Q weighted statistics

$$\hat{t}_q^{\text{BON}}(s; c_q) = \frac{(\mathbf{H}'_q \hat{\beta}_{ws} - c_q)^2}{\mathbf{H}'_q \hat{\mathbf{V}}_{mp}(\hat{\beta}_{ws}) \mathbf{H}_q}. \quad (5.2)$$

The largest statistic $\hat{t}_q^{\text{BON}}(s; c_q)$, for $q = 1, \dots, Q$, is compared to the upper tail of the distribution $F_{1, d}$ with a revised significance level α / Q instead of α . The number of degrees of freedom d is equal to the number of sampled primary sampling units minus the number of strata. Note that this procedure depends in general on the model parametrization used.

5.3 WALD F-procedure

An F-version of the standard Wald chi-square statistic, with adjusted denominator degrees of freedom as proposed by Fellegi (1980), can be defined as

$$\hat{t}^W(s; \mathbf{c}) = \frac{d - Q + 1}{Qd} (\mathbf{H} \hat{\beta}_{ws} - \mathbf{c})' (\mathbf{H} \hat{\mathbf{V}}_{mp}(\hat{\beta}_{ws}) \mathbf{H}')^{-1} (\mathbf{H} \hat{\beta}_{ws} - \mathbf{c}). \quad (5.3)$$

The statistic $\hat{t}^W(s; \mathbf{c})$ is compared to the upper tail of the distribution $F_{Q, d-Q+1}$. This procedure is implemented in the software package SUDAAN (Research Triangle Institute 2004).

5.4 Rao-Scott F-procedure

Another procedure consists of using an F-version (see Rao and Thomas 2003) of the second-order adjusted chi-square statistic of Rao and Scott (1981), which is based on

Satterthwaite's correction for the number of degrees of freedom. We use an adaptation of these authors' method for linear regression, as implemented in the software package SUDAAN (Research Triangle Institute 2004). The statistic is defined as

$$\hat{t}^{\text{RS}}(s; \mathbf{c}) = \frac{1}{\bar{\lambda} (1 + a^2) Q^*} (\mathbf{H} \hat{\beta}_{ws} - \mathbf{c})' (\mathbf{H} \hat{\mathbf{V}}_{\text{SRS}}(\hat{\beta}_{ws}) \mathbf{H}')^{-1} (\mathbf{H} \hat{\beta}_{ws} - \mathbf{c}), \quad (5.4)$$

where $\hat{\mathbf{V}}_{\text{SRS}}(\hat{\beta}_{ws})$ is an estimator of the variance-covariance matrix of $\hat{\beta}_{ws}$ under a simple random sampling design, $\bar{\lambda}$ is the average of the eigenvalues of the generalized design effect matrix $[\hat{\mathbf{V}}_{\text{SRS}}(\hat{\beta}_{ws})]^{-1} \hat{\mathbf{V}}_{mp}(\hat{\beta}_{ws})$, a is the coefficient of variation of these eigenvalues and $Q^* = Q / (1 + a^2)$. The Rao-Scott F-statistic $\hat{t}^{\text{RS}}(s; \mathbf{c})$ is compared to the upper tail of the distribution F_{Q^*, d^*} .

6. Simulation study

We performed a simulation study to investigate the level and power of the above test procedures in the case of informative and non-informative sampling. In sections 6.1 and 6.2, we describe the population and sample creation respectively. We then define the null hypotheses to be tested in section 6.3, describe the methods evaluated in section 6.4 and present simulation results in section 6.5.

6.1 Generation of the populations

We generated four populations of $N = 10,000$ units. First, a categorical variable v_k was generated independently for each population unit k so that $v_k = i$, for $i = 1, \dots, I$, with probability $P(v_k = i) = 1 / I$, where I is the number of categories of v_k , which was set equal to 5. The dependent variable y was generated as

$$y_k = \alpha_o + \alpha_1 \left(v_k - \frac{(I+1)}{2} \right) + \sigma \phi_k, \quad (6.1)$$

where $\phi_k \sim N(0, 1)$, $\alpha_o = 10$ and $\sigma = 3$. The four populations that we generated only differ in the choice of α_1 , which controls the correlation between y and v . We considered $\alpha_1 = 0, 0.25, 0.50$ and 0.75 .

6.2 Generation of samples and bootstrap weights

From each of the above four populations, 5,000 stratified simple random samples of size 100 were selected without replacement under two different stratification scenarios aimed at simulating both informative and non-informative sampling. In the case of non-informative sampling, the strata correspond exactly to the five categories of variable v defined above. In the case of informative sampling, the

strata are defined by the cross-classification of variable v and another categorical variable z that depends on the random error term $\sigma\phi_k$ in (6.1). For each population unit k , variable z was created as follows: $z_k = 1$, if $\sigma\phi_k > 0$, and $z_k = 2$, otherwise. This leads to 10 strata in the informative case that are constructed by crossing the five categories of v with the two categories of z . Each of the 10 informative strata contains about 1,000 population units while each of the 5 non-informative strata contains about 2,000 population units.

Furthermore, two different stratum allocation schemes were used. The scheme, SCHEME_UNEQUAL, allocates the 100 sample units among the strata in the following way:

Table 1
Sample sizes for SCHEME_UNEQUAL

Informative	$v \backslash z$	1	2	3	4	5
	1	4	4	16	4	28
	2	4	8	4	24	4
Non-informative		1	2	v	4	5
		8	12	20	28	32

The second scheme, denoted SCHEME_EQUAL, assigns the same number of units in each stratum as follows:

Table 2
Sample sizes for SCHEME_EQUAL

Informative	$v \backslash z$	1	2	3	4	5
	1	10	10	10	10	10
	2	10	10	10	10	10
Non-informative		1	2	v	4	5
		20	20	20	20	20

The two different schemes lead to very different sets of survey weights. The weights resulting from the SCHEME_UNEQUAL allocation are much more variable than those from SCHEME_EQUAL. Note that we simply defined the survey weight w_k as the inverse of the selection probability of unit k .

Finally, for each selected sample, 500 design-based bootstrap weights were calculated for each sampled unit, as described in Rao *et al.* (1992), among others. In our implementation of this methodology, each bootstrap sample was selected with replacement by stratified simple random sampling with $n_h - 1$ draws from the n_h sample units in stratum h . This methodology takes the sampling design variability into account (with a slight overestimation of the design variance due to assuming with-replacement sampling) but ignores the model variability. This is acceptable since the overall sampling fraction (1/100) is small.

6.3 Null hypotheses

For each selected sample, we modeled y_k as a function of v_k using an analysis of variance model. More specifically, we defined indicator variables

$$x_{ik} = \begin{cases} 1, & \text{if } v_k = i, \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \dots, I$, and fitted the linear model $y_k = \beta_0 + \sum_{i=1}^{I-1} \beta_i x_{ik} + \varepsilon_k$ using the weighted least-squares technique, where ε_k is a random error term with mean 0 and constant variance. We considered testing the following two null hypotheses:

$$\text{TEST1: } H_0: \beta_1 = 0$$

$$\text{TEST2: } H_0: \beta_1 = \beta_2 = \dots = \beta_{I-1} = 0.$$

Note that both null hypotheses are true for the population with $\alpha_i = 0$ while they are false for the other populations. The latter three populations are used to assess the power of the different test procedures under study.

6.4 Test methods

For each selected sample, we tested the above two null hypotheses using five different methods: the proposed bootstrap method, the naïve method (both unweighted and weighted versions) described in section 5.1, the Bonferroni method described in section 5.2, the Wald F method described in section 5.3 and the Rao-Scott F method described in section 5.4. Results for the naïve method are standard output in the software SAS whereas the Wald and Rao-Scott F-statistics are standard output in the SUDAAN statistical software, version 9. The Bonferroni statistics (5.2) are also obtained through SUDAAN. The proposed method is programmed in the statistical software SAS, version 8.

In addition, we also performed the simulation study using a linearized variance estimator in the Wald, Rao-Scott and Bonferroni methods instead of the bootstrap variance estimator (5.1). Rejection rates obtained using the linearized variance estimator were slightly lower but quite similar to those obtained using (5.1). Given this observation and that our focus is on bootstrap methods, we neither show nor discuss these additional results in the next section.

6.5 Simulation results

For each population, stratification scenario, allocation scheme, null hypothesis and method, we calculated the rejection rate in percentage over the 5,000 selected samples (using a 5% significance level). Results are given below in tables 3A, 3B, 4A and 4B. The results are more striking and more interesting for the null hypothesis TEST2 than the null

hypothesis TEST1. We will thus focus our discussion of the results on the former.

Tables 3A and 3B contain the results in the case of informative sampling, which is of more interest to us. Let us discuss first results in table 3A for SCHEME_UNEQUAL. Both naïve methods perform poorly as they do not properly exploit sampling design information. On the one hand, the unweighted version is definitely too liberal as its rejection rate is far above 5% under the null hypothesis. On the other hand, the weighted version is too conservative and significantly lacks power when compared to other methods. The Wald method is too liberal with a rejection rate of 15.8% when H_0 is true. The simple Bonferroni method improves the situation although it is still too liberal with a rejection rate of 11.4% when H_0 is true. This result is somewhat surprising as the Bonferroni method is known to be (asymptotically) conservative. A referee suggested that we consider an improved Bonferroni method such as that developed by Benjamini and Hochberg (1995). In this simulation study, such a method would not help as it always rejects more often than the standard Bonferroni method. The Rao-Scott method significantly outperforms the Wald and Bonferroni methods under the null hypothesis with a rejection rate of 6.8%. The proposed bootstrap method is comparable to the proven but more complicated Rao-Scott method with perhaps even a slight improvement in the level

with a rejection rate of 6.2% when H_0 is true. However, the Rao-Scott method is slightly more powerful than the proposed bootstrap method.

Table 3B contains results under SCHEME_EQUAL in the informative sampling scenario. Here, the weighted and unweighted versions of the naïve method yield similar results since the variability of the survey weights is quite small. Even in this case, the naïve method is definitely too conservative, which results in an extremely low power. All other methods are comparable both in terms of level (H_0 true) and power (H_0 false) although the Wald method is still slightly too liberal compared to the Bonferroni, Rao-Scott and proposed bootstrap methods with a rejection rate of 7.9% when H_0 is true.

Tables 4A and 4B contain the results in the case of non-informative sampling. Again, let us discuss first results in table 4A for SCHEME_UNEQUAL. As expected, the naïve unweighted method performs well here while the naïve weighted method becomes too liberal with a rejection rate of 12.8% when H_0 is true. In terms of the level, the proposed method is competitive to the naïve unweighted method and even slightly conservative. It outperforms the Wald method and is slightly better than the Bonferroni and Rao-Scott methods. Its power is however slightly less than these latter two competitors but still acceptable.

Table 3A
Rejection rates at the 5% significance level under SCHEME_UNEQUAL and informative sampling

SCHEME_UNEQUAL	Informative Sampling							
	Ho TRUE $\alpha_1 = 0$		Ho FALSE $\alpha_1 = 0.25$		Ho FALSE $\alpha_1 = 0.50$		Ho FALSE $\alpha_1 = 0.75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve Unweighted	37.5	100.0	85.3	100.0	98.8	100.0	100.0	100.0
Naïve Weighted	1.7	0.4	14.5	4.6	58.0	33.6	90.3	78.6
Wald	8.0	15.8	30.9	37.1	71.8	73.9	93.1	95.4
Rao-Scott	8.0	6.8	30.9	21.1	71.8	61.7	93.1	91.8
Bonferroni	8.0	11.4	30.9	32.6	71.8	68.8	93.1	91.9
Proposed Bootstrap	7.4	6.2	29.4	19.7	70.2	59.7	92.8	91.0

Table 3B
Rejection rates at the 5% significance level under SCHEME_EQUAL and informative sampling

SCHEME_EQUAL	Informative Sampling							
	Ho TRUE $\alpha_1 = 0$		Ho FALSE $\alpha_1 = 0.25$		Ho FALSE $\alpha_1 = 0.50$		Ho FALSE $\alpha_1 = 0.75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve Unweighted	0.1	0.0	6.7	0.3	58.1	16.5	97.2	79.7
Naïve Weighted	0.1	0.0	6.3	0.3	56.8	18.2	97.0	81.4
Wald	5.8	7.9	43.6	37.5	93.7	92.3	99.9	100.0
Rao-Scott	5.8	5.5	43.6	32.1	93.7	90.4	99.9	99.9
Bonferroni	5.8	6.2	43.6	33.6	93.7	88.6	99.9	99.8
Proposed Bootstrap	2.3	5.1	42.3	31.0	93.6	89.6	99.9	99.9

Table 4A
Rejection rates at the 5% significance level under SCHEME_UNEQUAL and non-informative sampling

Method	SCHEME_UNEQUAL		Non-Informative Sampling					
	Ho TRUE		Ho FALSE		Ho FALSE		Ho FALSE	
	$\alpha_1 = 0$		$\alpha_1 = 0.25$		$\alpha_1 = 0.50$		$\alpha_1 = 0.75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve Unweighted	4.2	4.7	13.5	11.2	39.9	34.6	71.8	70.5
Naïve Weighted	11.4	12.8	24.6	23.0	56.8	50.2	83.8	81.2
Wald	7.6	8.6	16.8	17.8	42.9	42.6	72.5	76.2
Rao-Scott	7.6	6.4	16.8	12.3	42.9	32.1	72.5	72.5
Bonferroni	7.6	7.1	16.8	16.5	42.9	42.1	72.5	75.0
Proposed Bootstrap	6.3	4.5	14.4	9.2	38.5	26.4	68.2	56.4

Table 4B
Rejection rates at the 5% significance level under SCHEME_EQUAL and non-informative sampling

Method	SCHEME_EQUAL		Non-Informative Sampling					
	Ho TRUE		Ho FALSE		Ho FALSE		Ho FALSE	
	$\alpha_1 = 0$		$\alpha_1 = 0.25$		$\alpha_1 = 0.50$		$\alpha_1 = 0.75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve Unweighted	4.9	4.5	17.2	12.4	54.3	42.2	88.2	81.7
Naïve Weighted	5.0	4.5	17.4	12.5	54.7	42.7	88.3	81.9
Wald	5.7	6.9	18.8	16.3	56.6	48.9	88.9	85.0
Rao-Scott	5.7	5.0	18.8	13.1	56.6	49.2	88.9	82.6
Bonferroni	5.7	5.4	18.8	13.7	56.6	45.1	88.9	81.8
Proposed Bootstrap	5.0	3.3	16.4	10.0	53.2	36.5	86.8	77.6

Table 4B contains results under SCHEME_EQUAL in the non-informative sampling scenario. In this table, the methods do not appear to differ drastically. As expected, the naïve method (both weighted and unweighted versions) performs well although it did not outperform the Rao-Scott and Bonferroni methods in this simulation study. The proposed method is still slightly conservative in this non-informative scenario and has slightly less power than the other methods.

To investigate the effect of large samples on the test procedures, we also performed some simulations with sample sizes that are ten times larger than in the original setup, as suggested by one reviewer. That is, we considered a population size of 100,000 and selected 1,000 samples of size 1,000 thus deliberately keeping the same small sampling fraction. From this setup, we obtained results when H_0 is true, shown in table 5, for both informative and non-informative sampling under unequal stratum allocation. As expected, all the methods other than the naïve ones have similar rejection rates that are indeed slightly lower than 5%. This illustrates that the differences between the methods become less important as the sample size increases.

Table 5
Rejection rates at the 5% significance level under SCHEME_UNEQUAL

Method	SCHEME_UNEQUAL		Informative		Non-informative	
			Ho TRUE		Ho TRUE	
			$\alpha_1 = 0$		$\alpha_1 = 0$	
	Test1	Test2	Test1	Test2	Test1	Test2
Naïve Unweighted	100.0	100.0	3.7	3.8		
Naïve Weighted	1.3	0.7	9.3	10.5		
Wald	4.6	4.5	3.2	4.1		
Rao-Scott	4.6	3.8	3.2	3.8		
Bonferroni	4.6	4.5	3.2	3.6		
Proposed Bootstrap	4.4	3.6	2.9	3.8		

Overall, our proposed bootstrap method was the best in terms of the level, followed closely by the Rao-Scott method. It gave somewhat conservative results in the non-informative sampling scenarios. This was accompanied by a slight loss of power. The Rao-Scott method is a good alternative if users have access to an appropriate software package. The Bonferroni method is simple to use but may be too liberal and the Wald method is even worse. The naïve methods may have serious deficiencies, either in the level or in the power, although the naïve unweighted method is viable if one is reasonably sure that sampling is not informative.

7. Summary and discussion

We have proposed a general and simple bootstrap procedure for testing hypotheses from survey data, which could also be applied outside the survey sampling field. Our procedure uses classical model-based test statistics and is thus easy to implement for analysts using classical software packages. We have shown in a simulation study that it performed well in the context of a linear regression model. These good results are encouraging and may suggest that our proposed bootstrap procedure could be useful with other more complicated models and other statistics. The idea could also be easily adapted for the construction of bootstrap confidence intervals.

One could also consider bootstrapping an asymptotically pivotal statistic such as the Rao-Scott statistic (5.4). This would however involve double bootstrapping if $\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$ is estimated using the bootstrap technique as in (5.1). Double bootstrapping requires generating another set of bootstrap replicates for each initial bootstrap replicate. Although better test procedures could potentially be obtained, double bootstrapping may not be convenient for analysts. By focusing on simpler statistics that do not involve the bootstrap technique, our test procedure avoids double bootstrapping and remains simple.

The properties of our method depend not only on the choice of the test statistic but also on the construction of the bootstrap weights. Typically, bootstrap weights capture the first two design moments of the sampling error, which should be sufficient in most cases to satisfy our bootstrap assumptions 3, 4 and 5. Bootstrap weights that also capture the third design moment could perhaps be useful for improving the level accuracy of the bootstrap test. This needs further investigation. Finally, as already pointed out in section 3, standard design-based bootstrap weights satisfy assumption 5 only when the overall sampling fraction is negligible so that the model portion of the total variance (3.1) is negligible. Research is needed to develop proper bootstrap weights, when a non-negligible sampling fraction is used, that capture both the model and the design portions of the total variance.

Acknowledgements

We sincerely thank the Associate Editor and two referees for their comments. We also thank J.N.K. Rao from Carleton University as well as David Binder and Yves Lafortune from Statistics Canada for their comments and stimulating discussions on this topic. All these comments and discussions were useful to improve the general quality of the paper and its clarity.

Appendix

Proof of result 1

From assumption 1, we can easily see that

$$\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} N(\mathbf{0}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'). \quad (\text{A.1})$$

Using a standard result on quadratic forms (e.g., Seber 1984, page 540) and equation (A.1), we obtain

$$n(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})' \hat{\mathbf{A}}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^Q \lambda_q \Omega_q, \quad (\text{A.2})$$

where λ_q , for $q = 1, \dots, Q$, are the eigenvalues of $\mathbf{A} = (\hat{\mathbf{A}}^{-1})(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$ and Ω_q are independent chi-square random variables with one degree of freedom. Therefore, from (A.2) and assumption 2, we have

$$\hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta}) =$$

$$(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})' \{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^Q \lambda_q \Omega_q.$$

References

- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.
- Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of survey data*, (Eds. R.L. Chambers and C.J. Skinner). New-York: John Wiley & Sons, Inc.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Efron, B., and Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.
- Graubard, B.I., and Korn, E.L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.
- Graubard, B.I., Korn, E.L. and Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 170-174.
- Hall, P., and Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757-762.

- Hughes, A.L., and Brodsky, M.D. (1994). Variance estimation of drug abuse episodes using the bootstrap. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 212-217.
- Kom, E.L., and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *The American Statistician*, 44, 270-276.
- Kom, E.L., and Graubard, B.I. (1991). A note on the large sample properties of linearization, jackknife and balanced repeated replication methods for stratified samples. *The Annals of Statistics*, 19, 2275-2279.
- Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, Supplement, 25-45.
- Langlet, É.R., Faucher, D. and Lesage, É. (2003). An application of the bootstrap variance estimation method to the Canadian Participation and Activity Limitation Survey. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2299-2306.
- Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- Rao, J.N.K., and Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: An appraisal and update. *Analysis of survey data*, (Eds. R.L. Chambers and C.J. Skinner). New-York: John Wiley & Sons, Inc.
- Research Triangle Institute (2004). *SUDAAN language manual, release 9.0*. Research Triangle Park, NC: Research Triangle Institute.
- Seber, G.A.F. (1984). *Multivariate Observations*. New-York: John Wiley & Sons, Inc.
- Shao, J., and Tu, D. (1995). *The jackknife and the bootstrap*. New-York: Springer-Verlag.
- Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, 20, 135-154.
- Thomas, D.R., and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.
- Thomas, D.R., Singh, A.C. and Roberts, G.R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64, 295-311.
- Yeo, D., Mantel, H. and Liu, T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye State) Polls

Bo-Seung Choi, Jai Won Choi and Yousung Park¹

Abstract

We use a Bayesian method to resolve the boundary solution problem of the maximum likelihood (ML) estimate in an incomplete two-way contingency table, using a loglinear model and Dirichlet priors. We compare five Dirichlet priors in estimating multinomial cell probabilities under nonignorable nonresponse. Three priors among them have been used for an incomplete one-way table, while the remaining two new priors are newly proposed to reflect the difference in the response patterns between respondents and the undecided. The Bayesian estimates with the previous three priors do not always perform better than ML estimates unlike previous studies, whereas the two new priors perform better than both the previous three priors and the ML estimates whenever a boundary solution occurs. We use four sets of data from the 1998 Ohio state polls to illustrate how to use and interpret estimation results for the elections. We use simulation studies to compare performance of the five Bayesian estimates under nonignorable nonresponse.

Key Words: Bayesian analysis; Nonignorable nonresponse; Contingency table; Boundary solution; EM algorithm.

1. Introduction

The problem of nonresponse is common in most surveys becoming a serious issue as the nonresponse rate increases (De Heer 1999; Groves and Couper 1998). When survey data is summarized in a two-way contingency table, the table includes fully classified counts, partially classified counts (*i.e.*, item nonresponse), and unclassified counts (*i.e.*, unit nonresponse). For example, in the Ohio (Buckeye State) Poll (BSP) (Chen and Stasny 2003), one category involves the voting preference (candidates A,B,C, or undecided) and the other category is the likelihood of voting (likely to vote, not likely to vote, and undecided). First supplemental margin contains data only on the voting preference, second contains data only on the likelihood of voting, and third is only the number of unit nonresponses (both responses unknown). Our interest is to incorporate these missing observations into estimating the true support for each candidate and to present Bayesian models to predict the winner.

In some surveys, the undecided answers are treated as a valid response category when the respondents do not have strong preference for a candidate and voting intention (Smith 1984; Rubin, Stern and Vehovar 1995). Many studies, however, have shown that the voting behavior of the undecided voters can have a significant impact on the final result and that by considering these undecided voters, the accuracy of election forecasting can be improved (Perry 1979; Fenwick, Wiseman, Becker and Heiman 1982; Myers and O'Connor 1983; Kim 1995; Chen and Stasny 2003; Martin, Traugott and Kennedy 2005). Perry (1979), among

them, showed that the undecided percentage in a poll is likely to be greater than the true percentage by presenting an empirical evidence using a secret ballot approach. Kim (1995) also indicated that these undecided voters are critical, especially in cases where the number of undecided voters is greater than the gap between the two leading runners in an election race. Three of our empirical studies in Section 3 belong to this critical case. Fenwick *et al.* (1982) and Kim (1995) applied a discriminant analysis to the October 1980 poll data in Massachusetts and the 1992 USA presidential election, from which they allocated the undecided voters to candidates to show that undecided voters generally do not vote in the same proportions as their decided counterparts. When the focus is on the candidate the undecided voter may vote for, undecided responses are better treated as missing data (Myers and O'Connor 1983). As indicated in Flannelly, Flannelly and McLeod (2000) and Lau (1994), the forecasting error for the actual election results increases as the rate of undecided voters increases. To overcome this problem, Monterola, Lim, Garcia and Saloma (2001) applied a neural network approach to classify undecided voters in a public opinion survey. Smith, Skinner and Clarke (1999) and Molenberghs, Kenward and Goetghebeur (2001) utilized model based imputation methods for the 1992 British General Election Panel Survey and the 1991 Slovenian plebiscite public opinion survey. Because our main goal is to obtain more accurate forecasts by allocating undecided voters to proper cell, we treat undecided voters as missing observations in the same way as these researchers handled them.

1. Bo-Seung Choi, Research Professor, Institute of Economics, Korea University, Seoul 136-701, Korea; Jai Won Choi, Professor, Department of Biostatistics, Medical College of Georgia, Augusta, GA 30912; Yousung Park, Professor, Department of Statistics, Korea University, Seoul 136-701, Korea. E-mail: yspark@korea.ac.kr.

Nonresponse (or undecided, equivalently) can be distinguished by three types of nonresponses (Little and Rubin 2002, page 11): missing completely at random (MCAR) means that the probability of a nonresponse on a variable of interest is independent of all survey variables including itself; missing at random (MAR) means that the probability of a nonresponse depends only on the observed data; missing not at random (MNAR) means that the probability of nonresponse depends on the unobserved values. Models for MCAR or MAR are called ignorable nonresponse models while models for MNAR are called nonignorable. For example, in a pre-election survey, if the respondents do not answer with their preference of a candidate, although they support a particular candidate, the pattern for candidate preference can be different between the respondents and nonrespondents. Then, the nonresponse mechanism is nonignorable. When data is assumed to be MCAR, the effect of nonresponse can be removed in likelihood inference (Little and Rubin 2002, page 11). However, when the nonrespondents follow a response pattern different from that of the respondents, discarding nonresponses or misspecifying the nonresponse mechanism leads to larger variances and biases in estimation (Chen 1972; Park and Brown 1994).

When nonresponse is nonignorable in contingency tables, ML estimation often yields boundary solutions where the probability of nonresponse is estimated to be zero in some cells. These boundary solutions often provide a local maximum of the likelihood function. In this case, the maximum likelihood (ML) estimates of the loglinear model parameters cannot have a unique solution and usually have large standard deviations (see Section 4 or Baker, Rosenberger and Dersimonian (1992) and Park and Brown (1994) for more detailed discussions).

The conditions where the ML estimate falls on the boundary solution have been proposed in a one-way contingency table (Baker and Laird 1988; Michiels and Molenbergs 1997). The geometric explanation for the boundary solution of the ML estimate was presented (Smith *et al.* 1999; Clark 2002). Baker *et al.* (1992) presented a sufficient and necessary condition under which the ML estimate can have a boundary solution in a two-way contingency table.

To overcome such a boundary problem in the ML estimate under the existence of nonignorable nonresponses, Park and Brown (1994) and Park (1998) proposed Bayesian approach using empirical priors based only on respondent information. Clogg, Rubin, Schenker and Schultz (1991) used constant prior for an incomplete one-way contingency table. Although they showed that, under nonignorable nonresponse, Bayesian methods provided smaller mean squared errors (MSE) than ML estimate in estimating cell

expectations, our simulation study shows that this is generally not true in an incomplete two-way contingency table. Thus, we present two Bayesian models whose priors depend on information from both respondents and undecided. We, then, apply each to analyze incomplete two-way contingency table. An extension to a multi-way table is straightforward. We can easily apply this extension to weighted data from stratified or cluster sampling using appropriate covariates (see Section 2.2).

The remainder of this paper is divided into four sections. In Section 2, we consider Bayesian models with five different priors and present a generalized expectation maximization (EM) algorithm to estimate cell probabilities. In Section 3, we apply the Bayesian models to four empirical data sets from the Buckeye State Poll (BSP) and compare the Bayesian estimates with the ML estimate and the actual election results. In Section 4, we use simulation studies to compare MSEs and biases of the Bayesian estimates from different missing percentages and response patterns of the respondents and nonrespondents. In this section, we also calculate the coverage probability to examine the performance of the Bayesian estimates. Section 5 includes some concluding remarks.

2. Bayesian models

We discuss five Bayesian estimates to accommodate nonignorable nonresponse in an incomplete two-way contingency table. We present an EM algorithm to tackle the nonresponse problem in a two-way contingency table in Section 2.1. Then, in Section 2.2, we specify five priors and extend our approach to a multi-way contingency table.

Let X_1 and X_2 be response variables indexed by I and J categories, respectively, in a two-way contingency table. We also let $R_1 = 1$ when X_1 is observed and $R_1 = 2$ when X_1 is missing. Similarly, $R_2 = 1$ when X_2 is observed and $R_2 = 2$ when X_2 is missing. Then the full array of X_1 , X_2 , R_1 , and R_2 constructs a $I \times J \times 2 \times 2$ contingency table in which we have completely classified counts, partially classified counts, and unclassified counts. To distinguish these three types of observations, let y_{ijkl} be the count belonging to the i^{th} category of X_1 , the j^{th} category of X_2 , the k^{th} value of R_1 , and the l^{th} value of R_2 . Thus, y_{ij11} is used for the completely classified counts, y_{i+12} and y_{+j21} for the respective column and row supplemental margins, and y_{++22} for the unclassified counts. We assume a multinomial distribution for these three types of observations to have the following log likelihood:

$$l = \sum_i \sum_j y_{ij11} \cdot \log(\pi_{ij11}) + \sum_i y_{i+12} \cdot \log(\pi_{i+12}) + \sum_j y_{+j21} \cdot \log(\pi_{+j21}) + y_{++22} \cdot \log(\pi_{++22}) \quad (1)$$

where $\pi_{ijkl} = \Pr[X_1 = i, X_2 = j, R_1 = k, R_2 = l]$ and $N = \sum_{i,j,k,l} \pi_{ijkl}$ is fixed.

Since this likelihood function involves more parameters than degrees of freedom available for estimation, we link π_{ijkl} to relevant covariates using a loglinear function. Since no explanatory variable is available, we do not use any explanatory variables. However, the loglinear model can easily incorporate explanatory variables in the same way as it incorporates the categorical variables (see Baker and Laird 1988 and Park and Brown 1994 for details).

A nonignorable nonresponse model for all of the variables X_1 , X_2 , R_1 , and R_2 is defined by

$$\log(m_{ijkl}) = \beta_0 + \beta_{X_1}^i + \beta_{X_2}^j + \beta_{R_1}^k + \beta_{R_2}^l + \beta_{X_1 R_1}^{ik} + \beta_{X_2 R_2}^{jl} + \beta_{X_1 X_2}^{ij} + \beta_{R_1 R_2}^{kl} \quad (2)$$

for $i = 1, \dots, I, j = 1, \dots, J, k = 1, 2,$ and $l = 1, 2$

where $m_{ijkl} = N \cdot \pi_{ijkl}$ is the expected cell count for the $(i, j, k, l)^{\text{th}}$ category and the sum of each β -term over any of its respective super script(s) is zero.

This loglinear model is saturated since the number of parameters is exactly the same as the number of cells observed from the incomplete two-way contingency table. This model is also a nonignorable nonresponse model because of the interaction terms between X_1 and R_1 and between X_2 and R_2 , implying that the nonresponse of each response variable depends on its own status. The loglinear model is a tool frequently used for analyzing incomplete contingency tables with nonignorable non-responses. Let p be the number of parameters (*i.e.*, β) to be estimated. We introduce the $p \times 1$ design vector \mathbf{z}_{ijkl} to indicate the affiliation of the observation belonging to the $(i, j, k, l)^{\text{th}}$ category. Then the loglinear model given in (2) can be rewritten as

$$\log \mathbf{m} = \mathbf{Z}\boldsymbol{\beta} \quad (3)$$

where the $I \times J \times 2 \times 2$ vector \mathbf{m} is the cell expectation and $\boldsymbol{\beta}$ is the vector representation of the β s. To avoid a boundary solution of the ML estimate in model (2), we impose Dirichlet priors to the cell probabilities $(\pi_{ij1}, \pi_{ij2}, \pi_{ij21}, \pi_{ij22})$ as given by

$$\prod_i \prod_j \pi_{ij11}^{\delta_{ij11}} \cdot \pi_{ij12}^{\delta_{ij12}} \cdot \pi_{ij21}^{\delta_{ij21}} \cdot \pi_{ij22}^{\delta_{ij22}} \quad (4)$$

where the hyper parameters, the δ_{ijkl} s are specified in Section 2.2. These Dirichlet priors produce an explicit and convenient form of a posterior distribution because they are conjugated to a multinomial distribution (Clogg *et al.* 1991; Park and Brown 1994; Forster and Smith 1998). Together with (3), the multinomial distribution of (1) for

observations, and the prior distribution (4), we have the following log posterior distribution:

$$\begin{aligned} l_{\text{pos}} = & \sum_i \sum_j y_{ij11} \cdot (\mathbf{z}_{ij11} \cdot \boldsymbol{\beta}) \\ & - \sum_i \sum_j y_{ij11} \cdot \log \left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right) \\ & + \sum_i y_{i+12} \cdot \log \left(\sum_j \exp(\mathbf{z}_{ij12} \cdot \boldsymbol{\beta}) \right) \\ & - \sum_i y_{i+12} \cdot \log \left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right) \\ & + \sum_j y_{+j21} \cdot \log \left(\sum_i \exp(\mathbf{z}_{ij21} \cdot \boldsymbol{\beta}) \right) \\ & - \sum_j y_{+j21} \cdot \log \left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right) \\ & + y_{++22} \cdot \log \left(\sum_i \sum_j \exp(\mathbf{z}_{ij22} \cdot \boldsymbol{\beta}) \right) \\ & - y_{++22} \cdot \log \left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right) \\ & + \sum_{i,j,k,l} \delta_{ijkl} \cdot (\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \\ & - \sum_{i,j,k,l} \delta_{ijkl} \cdot \log \left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right). \quad (5) \end{aligned}$$

Equation (5) is rather complex and thus we use the EM algorithm to estimate the parameters (*i.e.*, $\boldsymbol{\beta}$).

2.1 The EM algorithm

We maximize the posterior distribution given in (5) over the parameter $\boldsymbol{\beta}$ using the generalized expectation maximization (GEM) algorithm (Dempster, Laird and Rubin 1977) with the following E and M steps.

E-step: Using augmented y_{ij12} , y_{ij21} , and y_{ij22} for $i = 1, \dots, I$ and $j = 1, \dots, J$, the posterior (5) can be written as

$$\begin{aligned} l_{a,\text{pos}} = & \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \log(\pi_{ij11}) \\ & + \sum_i \sum_j (y_{ij12} + \delta_{ij12}) \log(\pi_{ij12}) \\ & + \sum_i \sum_j (y_{ij21} + \delta_{ij21}) \log(\pi_{ij21}) \\ & + \sum_i \sum_j (y_{ij22} + \delta_{ij22}) \log(\pi_{ij22}). \quad (6) \end{aligned}$$

To determine the expected augmented log posterior in (6), we average over the missing counts y_{ij12} , y_{ij21} , and y_{ij22} conditioning on the current parameter estimates, π_{ijkl}^{old} , and the marginal sums y_{+12} , y_{+j21} , and y_{++22} :

$$\begin{aligned}
E_{\text{old}}[I_{a,\text{pos}}] &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \cdot \log(\pi_{ij11}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij12} | \pi_{ijkl}^{\text{old}}, y_{i+12}] + \delta_{ij12}) \cdot \log(\pi_{ij12}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij21} | \pi_{ijkl}^{\text{old}}, y_{+j21}] + \delta_{ij21}) \cdot \log(\pi_{ij21}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij22} | \pi_{ijkl}^{\text{old}}, y_{++22}] + \delta_{ij22}) \cdot \log(\pi_{ij22}). \quad (7)
\end{aligned}$$

Since y_{ij12} , y_{ij21} , and y_{ij22} are multinomial random variates conditioned on the respective marginal sum y_{i+12} , y_{+j21} , and y_{++22} , the conditional expectations in the equation (7) are given by

$$\begin{aligned}
E_{\text{old}}(y_{ij12} | \pi_{ijkl}^{\text{old}}, y_{i+12}) &= y_{i+12} \frac{m_{ij12}^{\text{old}}}{m_{i+12}^{\text{old}}}, \\
E_{\text{old}}(y_{ij21} | \pi_{ijkl}^{\text{old}}, y_{+j21}) &= y_{+j21} \frac{m_{ij21}^{\text{old}}}{m_{+j21}^{\text{old}}},
\end{aligned}$$

and

$$E_{\text{old}}(y_{ij22} | \pi_{ijkl}^{\text{old}}, y_{++22}) = y_{++22} \frac{m_{ij22}^{\text{old}}}{m_{++22}^{\text{old}}}$$

where $m_{ijkl}^{\text{old}} = N \cdot \pi_{ijkl}^{\text{old}}$.

M-step: In this step, we maximize the expected log posterior (7) using the pseudo observations $\tilde{y}_{ij11} = y_{ij11} + \delta_{ij11}$, $\tilde{y}_{ij12} = y_{i+12} m_{ij12}^{\text{old}} / m_{i+12}^{\text{old}} + \delta_{ij12}$, $\tilde{y}_{ij21} = y_{+j21} m_{ij21}^{\text{old}} / m_{+j21}^{\text{old}} + \delta_{ij21}$, and $\tilde{y}_{ij22} = y_{++22} m_{ij22}^{\text{old}} / m_{++22}^{\text{old}} + \delta_{ij22}$. We impose the constraints on these pseudo observations so that their marginal sums are the same as the corresponding marginal sums of observations: $\tilde{y}_{++11} = y_{++11}$, $\tilde{y}_{i+12} = y_{i+12}$, $\tilde{y}_{+j21} = y_{+j21}$, and $\tilde{y}_{++22} = y_{++22}$. Under these constraints, the pseudo observations are now

$$\tilde{y}_{ijkl}^* = \begin{cases} \tilde{y}_{ij11} \frac{y_{++11}}{y_{++11} + \delta_{++11}} & \text{for } k = 1 \text{ and } l = 1 \\ \tilde{y}_{ij12} \frac{y_{i+12}}{y_{i+12} + \delta_{i+12}} & \text{for } k = 1 \text{ and } l = 2 \\ \tilde{y}_{ij21} \frac{y_{+j21}}{y_{+j21} + \delta_{+j21}} & \text{for } k = 2 \text{ and } l = 1 \\ \tilde{y}_{ij22} \frac{y_{++22}}{y_{++22} + \delta_{++22}} & \text{for } k = 2 \text{ and } l = 2. \end{cases}$$

Then, the expected log posterior function (7) becomes

$$\begin{aligned}
E_{\text{old}}[I_{a,\text{pos}}] &= \sum_i \sum_j \tilde{y}_{ij11}^* \cdot \log(\pi_{ij11}) \\
&+ \sum_i \sum_j \tilde{y}_{ij12}^* \cdot \log(\pi_{ij12}) \\
&+ \sum_i \sum_j \tilde{y}_{ij21}^* \cdot \log(\pi_{ij21}) \\
&+ \sum_i \sum_j \tilde{y}_{ij22}^* \cdot \log(\pi_{ij22}).
\end{aligned}$$

This equation has the same form as the likelihood obtained from a four-way contingency table with fully observed cell counts \tilde{y}_{ijkl}^* s. Thus, using the iterative re-weighted least squares method (Agresti 2002, page 342), we obtain the maximum posterior estimator (MPE) of β as follows:

$$\beta^{(r+1)} = (Z^T \hat{V}_t^{-1} Z)^{-1} Z^T \hat{V}_t^{-1} \gamma^{(r)},$$

where $\gamma^{(r)}$ has element $\gamma_{ijkl}^{(r)} = \log m_{ijkl}^{(r)} + (y_{ijkl} - m_{ijkl}^{(r)}) / m_{ijkl}^{(r)}$ and $\hat{V}_t = [\text{diag}(\mathbf{m}^{(r)})]^{-1}$. We finally iterate these E and M-steps until a convergence criterion is achieved. The convergence criterion we use is $\varepsilon \leq 10^{-6}$, where ε is the difference between two consecutive log posterior functions.

Let $Y_{\text{obs}} = (y_{ij11}, y_{i+12}, y_{+j21}, y_{++22})$ and $Y_{\text{mis}} = (y_{ij12}, y_{ij21}, y_{ij22})$ for $i = 1, \dots, I$ and $j = 1, \dots, J$ be the observed count vector and the missing count vector, respectively. Then the log posterior distribution (5) can be written as

$$\begin{aligned}
l_{\text{pos}} &= l(\beta | Y_{\text{obs}}) = l(\beta | Y_{\text{obs}}, Y_{\text{mis}}) \\
&- \log f(Y_{\text{mis}} | Y_{\text{obs}}, \beta). \quad (8)
\end{aligned}$$

By taking differentiation twice with respect to β , (8) yields

$$\begin{aligned}
\frac{\partial^2 l(\beta | Y_{\text{obs}})}{\partial \beta \partial \beta^T} &= \frac{\partial^2 l(\beta | Y_{\text{obs}}, Y_{\text{mis}})}{\partial \beta \partial \beta^T} \\
&- \frac{\partial^2 \log f(Y_{\text{mis}} | Y_{\text{obs}}, \beta)}{\partial \beta \partial \beta^T} \\
&= -Z^T [\text{diag}(\mathbf{m}) - \mathbf{m} \mathbf{m}^T / N] Z \\
&+ Z^T [\text{diag}(\pi) - \pi \pi^T] A B Z, \quad (9)
\end{aligned}$$

where π is vector expression of cell probabilities π_{ijkl} and A, B are given by

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \text{diag}\left(\frac{y_{i+12}^2}{y_{i+12} + \delta_{i+12}} \frac{m_{y12}}{m_{i+12}}\right) & 0 & 0 \\ 0 & 0 & \text{diag}\left(\frac{y_{+j21}^2}{y_{+j21} + \delta_{+j21}} \frac{m_{y21}}{m_{+j21}}\right) & 0 \\ 0 & 0 & 0 & \text{diag}\left(\frac{y_{+22}^2}{y_{+22} + \delta_{+22}} \frac{m_{y22}}{m_{+22}}\right) \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_{IJ} - B^{12} & 0 & 0 \\ 0 & 0 & I_{IJ} - B^{21} & 0 \\ 0 & 0 & 0 & I_{IJ} - B^{22} \end{pmatrix}$$

Here, to save the space and since there is no difficulty to extend for general i and j , B^{12} , B^{21} , and B^{22} are illustrated only for $I = 2$ and $J = 3$:

$$B^{12} = \begin{pmatrix} \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} \end{pmatrix},$$

$$B^{21} = \begin{pmatrix} \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \\ \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \end{pmatrix},$$

and

$$B^{22} = \begin{pmatrix} \frac{m_{1122}}{m_{+22}} & \frac{m_{1222}}{m_{+22}} & \frac{m_{1322}}{m_{+22}} & \frac{m_{2122}}{m_{+22}} & \frac{m_{2222}}{m_{+22}} & \frac{m_{2322}}{m_{+22}} \\ \frac{m_{1122}}{m_{+22}} & \frac{m_{1222}}{m_{+22}} & \frac{m_{1322}}{m_{+22}} & \frac{m_{2122}}{m_{+22}} & \frac{m_{2222}}{m_{+22}} & \frac{m_{2322}}{m_{+22}} \\ \frac{m_{1122}}{m_{+22}} & \frac{m_{1222}}{m_{+22}} & \frac{m_{1322}}{m_{+22}} & \frac{m_{2122}}{m_{+22}} & \frac{m_{2222}}{m_{+22}} & \frac{m_{2322}}{m_{+22}} \\ \frac{m_{1122}}{m_{+22}} & \frac{m_{1222}}{m_{+22}} & \frac{m_{1322}}{m_{+22}} & \frac{m_{2122}}{m_{+22}} & \frac{m_{2222}}{m_{+22}} & \frac{m_{2322}}{m_{+22}} \\ \frac{m_{1122}}{m_{+22}} & \frac{m_{1222}}{m_{+22}} & \frac{m_{1322}}{m_{+22}} & \frac{m_{2122}}{m_{+22}} & \frac{m_{2222}}{m_{+22}} & \frac{m_{2322}}{m_{+22}} \\ \frac{m_{1122}}{m_{+22}} & \frac{m_{1222}}{m_{+22}} & \frac{m_{1322}}{m_{+22}} & \frac{m_{2122}}{m_{+22}} & \frac{m_{2222}}{m_{+22}} & \frac{m_{2322}}{m_{+22}} \end{pmatrix}$$

We observe that the observed data information $\partial^2 l(\beta | Y_{\text{obs}}) / \partial \beta \partial \beta^T$ is equal to the augmented data information minus the missing data information. As shown in Gelman, Carlin, Stern and Rubin (2004, page 103), the inverse of the observed data information evaluated at the MPE of β is the variance of the MPE of β .

2.2 Specification of priors

To complete the EM algorithm, we need to determine the hyper-parameters, δ_{ijkl} s. We set the sum of priors $\sum_{i,j,k,l} \delta_{ijkl}$ equal to the number of parameters involved in the loglinear model, p , as suggested by Clogg *et al.* (1991). Under this constraint, we propose five types of priors as follows. We first allocate δ_{ijkl} for the MPE of m_{ijkl} to shrink toward the MLE obtained under ignorable non-response. That is, we determine δ_{ijkl} depending only on the known response counts y_{ij11} and call them respondent-driven priors.

The first type of respondent-driven prior is, for all $i = 1, \dots, I$ and $j = 1, \dots, J$,

$$\begin{aligned} \delta_{y11} &= \nabla_{11} \frac{y_{y11}}{y_{+11}}, \delta_{y12} = \nabla_{12} \frac{y_{y11}}{y_{+11}}, \delta_{y21} \\ &= \nabla_{21} \frac{y_{y11}}{y_{+11}}, \text{ and } \delta_{y22} = \nabla_{22} \frac{y_{y11}}{y_{+11}} \end{aligned} \quad (10)$$

where $\nabla_{kl} = p \cdot y_{++kl} / y_{++++}$ for $k = 1, 2$ and $l = 1, 2$. The second type of respondent-driven prior gives no prior (i.e., no need of prior as described below) on π_{y11} in the first type of priors. That is, the second type is the same as the first type except $\delta_{y11} = 0$ for all i and j . In the case of a one-way contingency table (i.e., either X_1 or X_2 is fully observed without missing information) and $y_{++22} = 0$, the first type is reduced to the priors used in Park (1998), whereas the second type is reduced to the priors used in Park and Brown (1994). These two types of respondent-driven priors may be too simplistic because the non-respondents are usually assumed to have a different response pattern from the respondents under a nonignorable nonresponse model. For example, the candidate preference of nonrespondents could be different from that of respondents in a pre-election survey.

In order to define the third type of prior, denote \hat{m}_{ijkl} as the MLE for m_{ijkl} . The closed form of \hat{m}_{ijkl} can be obtained from Baker *et al.* (1992) where some \hat{m}_{ijkl} could be zero because of boundary solutions. For example, when a supplemental column margin has a boundary solution in an incomplete 2×2 table, the MLEs are $\hat{m}_{1j11} = y_{1j11}$,

$$\hat{m}_{2j11} = \frac{y_{2+11}(y_{2j11} + y_{2j21})}{y_{2+11} + y_{++21}}, \hat{m}_{1j12} = \hat{m}_{1j11} b_j$$

where b_j is the solution of $\sum_{j=1}^2 y_{1j11} b_j = y_{1+12}$, $\hat{m}_{1j21} = 0$,

$$\hat{m}_{2j21} = \hat{m}_{2j11} \frac{y_{++21}}{y_{2+11}}, \hat{m}_{1j22} = 0,$$

and $\hat{m}_{2j22} = \hat{m}_{2j12} y_{++22} / y_{2+12}$. Therefore, these ML estimates accommodate both the information of respondents and nonrespondents, as well. The ML estimates can also be obtained from our EM algorithm in Section 2.1 by setting

$\delta_{ijkl} = 0$ for all i, j, k and l . Using these ML estimates, we define the third type of prior as

$$\begin{aligned} \delta_{y11} &= \nabla_{11} \cdot \left(\frac{\hat{m}_{y11}}{\hat{m}_{++11}} \right), \delta_{y12} \\ &= \nabla_{12} \cdot \left(\frac{\hat{m}_{y12}}{\hat{m}_{++12}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}, \\ \delta_{y21} &= \nabla_{21} \cdot \left(\frac{\hat{m}_{y21}}{\hat{m}_{++21}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}, \\ \text{and} \\ \delta_{y22} &= \nabla_{22} \cdot \left(\frac{\hat{m}_{y22}}{\hat{m}_{++22}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2} \end{aligned} \quad (11)$$

where $\nabla_{kl} = p \cdot \hat{m}_{++kl} / \hat{m}_{++++}$ for $k, l = 1, 2$, and the term $1/IJ$ is the constant prior of Clogg *et al.* (1991) to prevent possible boundary solutions for m_{y12} , m_{y21} , and m_{y22} (also see the fifth prior below). Thus, we allocate the third prior of δ_{ijkl} for the MPE of m_{ijkl} to shrink toward the ML obtained under the nonignorable nonresponse, whereas the first prior is obtained under an ignorable nonresponse model.

The fourth type of prior is defined by letting $\delta_{y11} = 0$ in (11) as we did in obtaining the second type of prior from the first type. The last type of prior is from Clogg *et al.* (1991) defined as

$$\delta_{y11} = 0, \delta_{y12} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right), \delta_{y21} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right), \text{and} \quad (12)$$

$$\delta_{y22} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right).$$

These five types of priors are summarized in Table 1 and are compared in the next section using empirical data and simulation studies.

Table 1

Five types of priors δ_{ijkl} (\hat{m}_{ijkl} is MLE, I and J are the numbers of row and columns in a two-way table, and p is the number of parameters)

	δ_{y11}	δ_{y12}	δ_{y21}	δ_{y22}	
Type I	$\nabla_{11} \frac{y_{y11}}{y_{++11}}$	$\nabla_{12} \frac{y_{y11}}{y_{++11}}$	$\nabla_{21} \frac{y_{y11}}{y_{++11}}$	$\nabla_{22} \frac{y_{y11}}{y_{++11}}$	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y_{++++}}$
Type II	0	$\nabla_{12} \frac{y_{y11}}{y_{++11}}$	$\nabla_{21} \frac{y_{y11}}{y_{++11}}$	$\nabla_{22} \frac{y_{y11}}{y_{++11}}$	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y_{++++}}$
Type III	$\nabla_{11} \cdot \left(\frac{\hat{m}_{y11}}{\hat{m}_{++11}} \right)$	$\nabla_{12} \cdot \left(\frac{\hat{m}_{y12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\nabla_{21} \cdot \left(\frac{\hat{m}_{y21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\nabla_{22} \cdot \left(\frac{\hat{m}_{y22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right)$	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}_{++++}}$
Type IV	0	$\nabla_{12} \cdot \left(\frac{\hat{m}_{y12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\nabla_{21} \cdot \left(\frac{\hat{m}_{y21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\nabla_{22} \cdot \left(\frac{\hat{m}_{y22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right)$	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}_{++++}}$
Type V	0	$\nabla_{12} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{21} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{22} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{kl} = \frac{p}{3}$

$$y_{++++}^* = y_{++++} - y_{++11} \text{ and } \hat{m}_{++++}^* = \hat{m}_{++++} - \hat{m}_{++11}$$

Up to this point, we have presented methods for a two-way table, and y_{ijk} is defined for the count of the (i, j) cell of the i^{th} row and j^{th} column (i.e., $X_1 = i, X_2 = j$), and indicator R_1 for a missing row and R_2 for a missing column (i.e., $R_1 = k, R_2 = l$). This can be easily extended to the 3-way table. Denote y_{ijklmn} to be the $(i, j, k)^{\text{th}}$ cell count for the three response variables (i.e., $X_1 = i, X_2 = j$, and $X_3 = k$) and respective missing rows and columns (i.e., $R_1 = l, R_2 = m$, and $R_3 = n$ for $l, m, n = 1, 2$). Thus, $lmn = 111$ implies that all of the three variables are observed, $lmn = 112$ implies that X_1 and X_2 are observed but X_3 is missing; similarly for $lmn = 121, 122, 211, 212, 221, 222$; 1 is for observed and 2 designates missing. Accordingly, the EM algorithm and priors for an incomplete three-way contingency table can be defined. The conditional expectation in the E-step for the $(i, j, k)^{\text{th}}$ cell with unknown information of the k margin is

$$E_{\text{old}}(y_{ijk112} | \pi_{ijklmn}^{\text{old}}, y_{ij+112}) = y_{ij+112} \frac{m_{ijk112}^{\text{old}}}{m_{i+112}^{\text{old}}}.$$

Similarly,

$$E_{\text{old}}(y_{ijk122} | \pi_{ijklmn}^{\text{old}}, y_{i++122}) = y_{i++122} \frac{m_{ijk122}^{\text{old}}}{m_{i++122}^{\text{old}}}.$$

and

$$E_{\text{old}}(y_{ijk222} | \pi_{ijklmn}^{\text{old}}, y_{+++222}) = y_{+++222} \frac{m_{ijk222}^{\text{old}}}{m_{+++222}^{\text{old}}}.$$

Other expectations and five types of priors can be similarly defined.

The Buckeye state poll is a Random Digit Dialing (RDD). No modification is necessary for the Bayesian procedures if the RDD is strictly a self-weighting survey (Lavrakas 1993; Potthoff 1994). However, RDD is not always done by a self-weighting design. For example, a telephone sample comprises a sample of households, not persons. If one person is interviewed in a household, a weight should be superimposed on the response by the number of persons in the household. A weight is also needed for the households with more than one telephone number. If an accurate estimate of the total number of households is available, stratification by region or state is possible and weighting must be considered in a comprehensive analysis. RDD was used in the 1998 Ohio election surveys. In this study, our method and models do not include weighting from stratification, clustering, and other factors leading to different probabilities of selection in a telephone survey.

However, further extension can be made for such weighting. A simple extension below shows how to accommodate a typical stratification. In a three-way table, let X_3 be the third response variable indexed by h

($h = 1, \dots, H$) that is assumed to be always observed. The H categories can be strata in a stratified sampling. Since X_3 is always observed, the corresponding missingness variable R_3 is equal to 1 and its observation can be denoted by y_{ijk1m} . Then, we can write the following log likelihood for each stratum h :

$$l_h = \sum_{i=1}^I \sum_{j=1}^J y_{ijh11} \log(\pi_{ijh11}) + \sum_{i=1}^I y_{i+h121} \log(\pi_{i+h121}) \\ + \sum_{j=1}^J y_{+jh211} \log(\pi_{+jh211}) + y_{++h221} \log(\pi_{++h221})$$

where $\pi_{ijhlm} = P[X_1 = i, X_2 = j, R_1 = l, R_2 = m | X_3 = h]$. Thus, the terminology X_3 used for a three-way table acts as an indicator for strata. For each stratum h , the likelihood of (13) is exactly the same as that of a 2-way table.

Then, a log linear model for the cell expectation $m_{ijhlm} = N_h \cdot \pi_{ijhlm}$ can be defined in a similar way as in (2) where $N_h = \sum_{i,j,l,m} y_{ijhlm}$ for each $h = 1, 2, \dots, H$. A nonignorable nonresponse model is given by

$$\log(m_{ijhlm}) = \beta_{0h} + \beta'_{X_1h} + \beta'_{X_2h} + \beta'_{R_1h} \\ + \beta'_{R_2h} + \beta'_{X_1X_2h} + \beta'_{X_1R_1h} + \beta'_{X_2R_2h} \quad (13)$$

To avoid a boundary solution problem as in Section 2, we use the Dirichlet priors for π_{ijhlm}

$$\prod_i \prod_j \pi_{ijh11}^{\delta_{ijh11}} \cdot \pi_{ijh12}^{\delta_{ijh12}} \cdot \pi_{ijh21}^{\delta_{ijh21}} \cdot \pi_{ijh22}^{\delta_{ijh22}}.$$

Then, we follow exactly the same procedures as shown in Section 2 to estimate the cell expectations m_{ijhlm} for each $h = 1, 2, \dots, H$. The estimate of the $(i, j)^{\text{th}}$ cell expectation is

$$\hat{E}(y_{ij}) = \sum_{h=1}^H w_h \sum_{l,m} \hat{m}_{ijhlm}$$

where w_h is the known weight for the h^{th} stratum and \hat{m}_{ijhlm} is the m_{ijhlm} evaluated at the MPE of β . For example, $w_h = N_h / \sum_h N_h$ is for a stratified sample where N_h is the population size of the h^{th} stratum.

The variance-covariance matrix of an approximation to the distribution of $\hat{\mathbf{m}}$ is

$$\frac{\partial \hat{\mathbf{m}}}{\partial \beta}^T \text{Var}(\hat{\beta}_{\text{MPE}}) \frac{\partial \hat{\mathbf{m}}}{\partial \beta} \quad (14)$$

where $\hat{\mathbf{m}}$ is a vector expression of the cell estimates \hat{m}_{ijhlm} , $\hat{\beta}_{\text{MPE}}$ is the MPE of β and its variance $\text{Var}(\hat{\beta}_{\text{MPE}})$ is given by the inverse of (9), and $\partial \mathbf{m} / \partial \beta = N_h \times [\text{diag}(\hat{\pi}) - \hat{\pi} \hat{\pi}^T] \mathbf{Z}$ where $\hat{\pi}$ has

$$\hat{\pi}_{ijhlm} = \pi_{ijhlm}(\hat{\mathbf{p}}_{\text{MPE}}) = \frac{\exp(z_{ijhlm}\hat{\mathbf{p}}_{\text{MPE}})}{\sum_{k \in (i,j,h,l,m)} \exp(z_k\hat{\mathbf{p}}_{\text{MPE}})}$$

as its typical element.

3. An application to a Buckeye State Poll

In forecasting the winner in a poll, the accuracy of the poll often depends on how to handle undecided voters who are likely to vote but who have not yet decided their preference for a candidate. We compare the Bayesian estimates based on the five types of priors with the ML estimate using the Buckeye State Poll (BSP) conducted in 1998 by the Center for Survey Research at Ohio State University. The BSP surveys produced incomplete two-way contingency tables with one category being candidate preference and the other category being the likelihood of voting in the November 1998 races for Ohio Governor, Attorney-General, Mayor of Columbus, and Treasurer. Table 2 summarizes these four polls and shows a substantial number of undecided voters.

For comparison, we consider the following ignorable Model 1 and the two nonignorable nonresponse Model 2 and Model 3.

$$\text{Model 1: } \log(m_{ijkl}) = \beta_0 + \beta_{X_1}^i + \beta_{X_2}^j + \beta_{R_1}^k + \beta_{R_2}^l + \beta_{X_1X_2}^{ij} + \beta_{R_1R_2}^{kl},$$

$$\text{Model 2: } \log(m_{ijkl}) = \beta_0 + \beta_{X_1}^i + \beta_{X_2}^j + \beta_{R_1}^k + \beta_{R_2}^l + \beta_{X_1R_1}^{ik} + \beta_{X_2R_2}^{jl} + \beta_{X_1X_2}^{ij} + \beta_{R_1R_2}^{kl},$$

$$\text{Model 3: } \log(m_{ijkl}) = \beta_0 + \beta_{X_1}^i + \beta_{X_2}^j + \beta_{R_1}^k + \beta_{R_2}^l + \beta_{X_1R_2}^{il} + \beta_{X_2R_1}^{jk} + \beta_{X_1X_2}^{ij} + \beta_{R_1R_2}^{kl}.$$

Model 1 is missing completely at random, and cases with missing data can be ignorable in likelihood inferences. Model 2 and Model 3 are nonignorable where the probability of missing a variable depends on itself in Model 2 while the probability in Model 3 depends on the other variable. Note that the ML estimates in Model 1 and Model 3 are not on the boundary of the parameter space as shown by Baker *et al.* (1992). Moreover, since we found that, under Model 1 and Model 3, all of the five Bayesian estimates for the expected cell counts are not only fairly close to the ML estimate and their standard deviations are almost the same, we only present the ML estimates for Model 1 and Model 3.

We denote the ML estimates under ignorable Model 1, nonignorable Model 2, and nonignorable Model 3 by $IG1_{ML}$, $NON2_{ML}$, and $NON3_{ML}$, respectively. IG and NON stand for ignorable and nonignorable, respectively. We also let $NON2_i^{BE}$ be the Bayesian estimator using the i^{th} type of priors under Model 2. That is, $NON2_1^{BE}$ uses the respondent-driven priors of (10) and $NON2_2^{BE}$ is the same priors as $NON2_1^{BE}$ except for $\delta_{ij11} = 0$. Similarly, $NON2_3^{BE}$ is given by (11) and $NON2_4^{BE}$ is the same priors except for $\delta_{ij11} = 0$. $NON2_5^{BE}$ is the Bayesian estimate using the constant priors of (12). In addition, we can use the Stasny method (1986, 1988) to estimate the expected cell counts under Model 1 and Model 3 that she implicitly assumed. However, her estimates appear to be exactly the same as $IG1_{ML}$.

Table 2
Observed data for BSP pre-election surveys

	Governor race				Attorney-general race		
	Fisher	Taft	Others	Undecided	Montgomery	Cordray	Undecided
Likely to vote	112	140	23	61	197	82	57
Unlikely to vote	96	108	21	73	161	65	75
Undecided	7	11	1	4	15	4	0
	Mayor race				Treasurer race		
	Coleman	Teater	Espy	Undecided	Deters	Donofrio	Undecided
Likely to vote	40	32	25	30	127	119	90
Unlikely to vote	37	47	41	56	127	90	84
Undecided	0	2	1	0	10	7	0

The top table in Table 3 shows predicted values of elections using only “likely to vote” for the four races and their standard deviations in parentheses. The standard deviations are close to each other and show significant differences between the first and second leading candidates, except in the race for Mayor. This table also includes the actual election results and shows whether or not the ML estimates fall into the boundary solutions.

The bottom table shows the predictions of elections using both “likely to vote” and “unlikely to vote” to see what happens if those who responded to “unlikely to vote” actually voted. Comparing the two tables, we may conclude that the winners for Governor, Attorney-General, and the Treasurer’s elections remained unchanged regardless of the likelihood of voting, whereas the winner could have changed in the Mayor’s election if most of those who were “unlikely to vote” actually voted.

Based on Table 3, we can classify the 7 estimates, except $NON2_{ML}$, into two groups: $NON2_3^{BE}$, $NON2_4^{BE}$, and $NON2_5^{BE}$ to the first group and the remaining four estimates, $NON2_1^{BE}$, $NON2_2^{BE}$, $IG1_{ML}$, and $NON3_{ML}$ to

the second group. As expected, since the priors δ_{ijkl} for $NON2_1^{BE}$ and $NON2_2^{BE}$ are so defined that the estimate of m_{ijkl} shrinks toward the ML under an ignorable nonresponse model, these two Bayesian estimates are very close to $IG1_{ML}$ and hence have little advantage over the $IG1_{ML}$. It is also interesting to note that $NON3_{ML}$ is almost the same as $IG1_{ML}$ although their loglinear models are differently specified.

There is no general criterion to evaluate whether an ignorable nonresponse model or a nonignorable non-response model is appropriate. However, as stated in Chen and Stasny (2003), the assumption of nonignorability for a nonresponse may be a reasonable assumption in the Buckeye State Poll study because people might be reluctant to express their preference for an unpopular candidate, or their current preferences are not firm or accurate at the time of the interview. In this regard, the $NON2_1^{BE}$, $NON2_2^{BE}$, and $NON3_{ML}$ may not be appropriate in these particular case studies because they are almost the same as the $IG1_{ML}$ of Model 1.

Table 3
Prediction of elections based on the October 98 and April 98 Buckeye State Polls (the unit is % and the numbers in parentheses are standard deviations)

	Governor			Mayor			Attorney-General		Treasurer	
	Fisher	Taft	Others	Coleman	Teater	Espy	Mongomery	Cordray	Deters	Donofrio
	Likely to vote only used									
$NON2_{ML}$	33.2(2.75)	42.1(3.00)	24.8	31.5(4.65)	25.3(4.23)	43.2	75.6(3.71)	24.4	57.0(3.48)	43.0
$NON2_1^{BE}$	40.6(3.04)	48.5(3.27)	10.9	38.1(5.14)	34.2(4.78)	27.7	72.1(3.61)	27.9	52.7(3.36)	47.3
$NON2_2^{BE}$	40.9(3.01)	50.7(3.20)	8.40	39.9(5.04)	33.6(4.83)	26.5	71.0(3.59)	29.0	52.1(3.34)	47.9
$NON2_3^{BE}$	35.8(2.85)	44.5(3.08)	19.7	35.6(4.87)	29.3(4.51)	35.1	63.0(3.67)	37.0	54.3(3.41)	45.7
$NON2_4^{BE}$	36.3(2.87)	45.2(3.11)	18.6	35.9(4.91)	29.4(4.52)	34.6	63.0(3.64)	37.0	53.9(3.40)	46.1
$NON2_5^{BE}$	38.9(2.99)	47.4(3.20)	13.7	37.7(4.99)	33.6(4.77)	28.7	66.0(3.54)	34.0	51.5(3.32)	48.5
$IG1_{ML}$	40.6(3.03)	51.2(3.28)	8.20	40.8(5.16)	33.4(4.76)	25.8	70.9(3.59)	29.1	51.8(3.32)	48.2
$NON3_{ML}$	40.6(3.03)	51.2(3.28)	8.20	40.9(5.16)	33.3(4.75)	25.8	70.9(3.58)	29.1	51.7(3.32)	48.3
Actual result	45	50	5	39	37	24	63	37	57	43
Boundary	yes			yes			yes		no	
	Likely to vote + Unlikely to vote									
$NON2_{ML}$	32.7(1.83)	39.4(1.91)	27.8	24.8(2.45)	26.2(2.49)	49.0	77.0(1.64)	23.0	60.2(1.93)	39.8
$NON2_1^{BE}$	41.3(1.93)	46.4(1.96)	12.3	30.7(2.68)	37.1(2.75)	32.2	72.8(1.74)	27.2	56.0(1.96)	44.0
$NON2_2^{BE}$	41.9(1.93)	49.2(1.95)	8.90	32.7(2.63)	36.5(2.76)	30.8	71.4(1.77)	28.6	55.3(1.96)	44.7
$NON2_3^{BE}$	35.4(1.87)	41.8(1.93)	22.7	27.8(2.55)	30.5(2.62)	41.7	61.0(1.72)	39.0	57.6(1.95)	42.4
$NON2_4^{BE}$	36.0(1.88)	42.6(1.93)	21.4	28.7(2.57)	30.6(2.62)	40.7	60.9(1.75)	39.1	57.2(1.95)	42.8
$NON2_5^{BE}$	39.1(1.91)	45.1(1.95)	15.8	30.7(2.63)	35.8(2.74)	33.5	64.8(1.88)	35.2	54.8(1.96)	45.2
$IG1_{ML}$	41.5(1.96)	49.8(1.96)	8.70	33.9(2.70)	36.1(2.74)	29.9	71.2(1.78)	28.8	55.0(1.96)	45.0
$NON3_{ML}$	41.5(1.96)	49.8(1.96)	8.70	34.1(2.71)	36.0(2.74)	29.9	71.1(1.78)	28.9	55.0(1.96)	45.0

Compared to actual election results, $NON2_{ML}$ gives the worst prediction for Governor, Mayor, and Attorney-General because the $NON2_{ML}$ lies on a boundary solution; whereas it provides the best prediction for Treasurer because it does not lie on a boundary solution. In the Attorney-General's election, $NON2_3^{BE}$ and $NON2_4^{BE}$ not only predicted the exact actual result but also are quite different from the other estimates. Since $NON2_3^{BE}$ and $NON2_4^{BE}$ have the priors to reflect different response patterns between respondents and the undecided, we can infer that the undecided voters in the Attorney-General race have quite different preference for the candidate from the respondents (*i.e.*, $NON2_3^{BE}$ and $NON2_4^{BE}$ allocate 19.4 % of the undecided voters who are likely to vote for Montgomery and 80.6% for Cordray, whereas the data in Table 2 indicates the percentage of Montgomery vs Cordray is 29.4% vs 70.6% among respondents who are likely to vote).

To see this difference between the respondents and undecided voters in terms of parameter estimates and to examine the effect of occurrence of the boundary solution on the estimates under the nonignorable Model 2, we present the ML estimates and $NON2_3^{BE}$ estimates and their corresponding standard deviations for the Attorney-General race in Table 4. Because of a boundary solution, all of the ML estimates have too large standard deviations as expected. On the other hand, $NON2_3^{BE}$ is very stable. Since $\beta_{x_1x_2}^{11} = 0.0472$ is the smallest and its standard deviation is relatively large, we neglect $\beta_{x_1x_2}^{11}$ to avoid complexity of interpretation. Under $\beta_{x_1x_2}^{11} = 0$, it is not difficult to show that, using the estimates of $NON2_3^{BE}$ in Table 4,

$$\log \frac{m_{1j|l}}{m_{2j|l}} = 2(\beta_{x_1}^1 + \beta_{x_1r_1}^{11}) = 0.09$$

and

$$\log \frac{m_{1j|2l}}{m_{2j|2l}} = 2(\beta_{x_1}^1 - \beta_{x_1r_1}^{11}) = 1.3916$$

for each fixed j and l , and

$$\log \frac{m_{ikl}}{m_{i2k1}} = 2(\beta_{x_2}^1 + \beta_{x_2r_2}^{11}) = 0.8982$$

and

$$\log \frac{m_{ik2}}{m_{i2k2}} = 2(\beta_{x_2}^1 - \beta_{x_2r_2}^{11}) = -1.4942$$

for each fixed i and k . Thus, by

$$\log \frac{m_{1j|l}}{m_{2j|l}} = 2(\beta_{x_1}^1 + \beta_{x_1r_1}^{11}) = 0.09,$$

those who are likely to vote (*i.e.*, $i = 1$) are 1.09 times (*i.e.*, $e^{0.09}$) more than those who are unlikely to vote (*i.e.*, $i = 2$) among respondents ($k = 1$), whereas, by

$$\log \frac{m_{1j|2l}}{m_{2j|2l}} = 2(\beta_{x_1}^1 - \beta_{x_1r_1}^{11}) = 1.3916,$$

likely voters of $i = 1$ are 4.02 times (*i.e.*, $e^{1.3916}$) more than unlikely voters of $i = 2$ among undecided ($k = 2$); by

$$\log \frac{m_{ikl}}{m_{i2k1}} = 2(\beta_{x_2}^1 + \beta_{x_2r_2}^{11}) = 0.8982,$$

those who vote for Montgomery are 2.46 times more than those who vote for Cordray among respondents; whereas, by

$$\log \frac{m_{ik2}}{m_{i2k2}} = 2(\beta_{x_2}^1 - \beta_{x_2r_2}^{11}) = -1.4942,$$

unlikely voters are 4.46 times more than likely voters among the undecided. This implies that the response pattern is much different between respondents and the undecided.

Table 4
ML and the third type Bayesian Estimates under nonignorable Model 2 for Attorney-General (the standard deviations are in parentheses)

	β_0	$\beta_{x_1}^1$	$\beta_{x_2}^1$	$\beta_{r_1}^1$	$\beta_{r_2}^1$	$\beta_{x_1r_1}^{11}$	$\beta_{x_2r_2}^{11}$	$\beta_{x_1x_2}^{11}$	$\beta_{r_1r_2}^{11}$
$NON2_{ML}$	-3.3735	-1.9487 (3.120)	3.2134 (8.515)	4.8496 (3.996)	4.8186 (8.871)	2.0283 (3.120)	-2.7594 (8.512)	-0.0452 (0.045)	-1.5588 (2.501)
$NON2_3^{BE}$	0.6860	0.3704 (0.118)	-0.1490 (0.052)	3.3024 (2.501)	2.2942 (2.501)	-0.3254 (0.117)	0.5981 (0.052)	0.0472 (0.041)	-1.5450 (2.501)

The extent of this difference can be measured by the most important terms, $\beta_{X_1 R_1}^{11}$ and $\beta_{X_2 R_2}^{11}$, in the nonignorable nonresponse model, Model 2. Since

$$\beta_{X_1 R_1}^{11} = \frac{1}{4} \log \frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = -0.3254$$

and

$$\beta_{X_2 R_2}^{11} = \frac{1}{4} \log \frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = 0.5981, \beta_{X_1 R_1}^{11}$$

is the log-odds ratio that shows the log difference between the ratio of the number of those “likely to vote” to that of those “unlikely to vote” among the decided voters for Montgomery and the same ratio among the undecided voters who prefer Montgomery but who do not express their likelihood of voting. Whereas, $\beta_{X_2 R_2}^{11}$ is the log-odds ratio that shows the log difference between the ratio of the number of voters for Montgomery to the voters for Cordray among the decided who are likely to vote and the same ratio among the undecided voters who are likely to vote but who do not express their candidate preference. Thus, among voters for Montgomery, the possibility for the undecided voters to vote relative to not voting is about 3.67 times

$$\left(\text{i.e., } \frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = e^{4 \times -0.3254} = 3.67^{-1} \right)$$

larger than the possibility of the decided, implying that Montgomery needs a strategy to raise the turnout of voters. On the other hand, among those likely to vote, the supporting rate of the decided for Montgomery is about 10.94 times

$$\left(\text{i.e., } \frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = e^{4 \times 0.5981} = 10.94 \right)$$

larger than the undecided voters for Montgomery, implying that most of the undecided voters not exposing their preference of candidate are likely to vote for Cordray as the Attorney-General. This also confirms the popular account that voters are inclined to remain “undecided” in a poll if they support the candidate who is seen as inferior in a race and that the voters are inclined to abstain from voting if they support the candidate who certainly dominates the race.

4. Simulation study

We consider a 2×2 contingency table with supplemental margins to compare the performance of the five Bayesian estimates described in Section 2 for different missing percentages and different response patterns under the following nonignorable nonresponse model (i.e., Model 2):

$$\begin{aligned} \log(m_{ijkl}) = & \beta_0 + \beta_{X_1}^i + \beta_{X_2}^j + \beta_{R_1}^k + \beta_{R_2}^l \\ & + \beta_{X_1 R_1}^{ik} + \beta_{X_2 R_2}^{jl} + \beta_{X_1 X_2}^{ij} + \beta_{R_1 R_2}^{kl}. \end{aligned}$$

Thus, we only compare $NON2_{ML}$ and $NON2_{X_i}^{BE}$ for $i = 1, \dots, 5$ in this simulation study.

Since there are two levels in all of X_1 , X_2 , R_1 , and R_2 , there are 8 parameters to be determined for the simulation study. From the equations of

$$4\beta_{X_1 R_1}^{11} = \log \frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} \text{ and } 4\beta_{X_2 R_2}^{11} = \log \frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}},$$

$$\beta_{X_1 R_1}^{11} = \beta_{X_2 R_2}^{11} = 0$$

means that there is no difference in the response pattern between respondents and undecided. The bigger $\beta_{X_1 R_1}^{11}$ and $\beta_{X_2 R_2}^{11}$ are, the more different the response pattern between respondents and undecided voters is. We vary these two parameters from 0.2 to 0.8 with an increment of 0.2. We set the missing percentage to 20% and 30% by adjusting $\beta_{X_1}^1$ and $\beta_{R_1}^1$ and fixing

$$\frac{m_{1111}/m_{2111}}{m_{2111}/m_{2211}} = 5, \frac{m_{1111}/m_{1112}}{m_{1112}/m_{1122}} = 2,$$

and

$$N = \sum_{ijkl} m_{ijkl} = 1,000.$$

This implies that the size and missing percentage for the cell of $X_1 = 1$ and $X_2 = 1$ are approximately 5 times and 2 times the size of the other three cells, respectively.

We generate a large number of samples $\{y_{ijkl}, i, j, k, l = 1, 2\}$ from the above setting until we have 1,000 random samples with boundary solutions and the other 1,000 with no boundary solutions. The occurrence of a boundary solution is determined by the criterion given in Michiels and Molenberghs (1997) (also see Clarke (2002), Smith *et al.* (1999) for more details). Using $\{y_{ij11}, y_{i+12}, y_{+j21}, y_{++22}, i, j, = 1, 2\}$ obtained from the generated data, the expected cell counts m_{ijkl} ’s are estimated by each of the five Bayesian estimates and the ML estimate described in Section 2.

We calculate mean squared errors (MSEs) and absolute biases of $NON2_{ML}$, $NON2_1^{BE}$, ..., $NON2_5^{BE}$ for $\{\sum_{kl} m_{ijkl}, i, j = 1, 2\}$. Then we take the mean over the four MSEs and the four absolute biases, which we obtain from each estimate to see the overall performance of the estimate. Similarly, we calculate mean MSEs and mean absolute biases for $\{m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$ to see the performance of each estimate in imputing the nonresponses.

Table 5 shows the ratios of the mean MSEs and mean absolute biases of the five Bayesian estimates (*i.e.*, $NON2_1^{BE}$, ..., $NON2_5^{BE}$), relative to the ML estimate (*i.e.*, $NON2_{ML}$) when the boundary solutions occur; whereas Table 5 shows the ratios when no boundary occurs. Thus, values less than 1 imply that the corresponding Bayesian estimate has a smaller mean MSE or a smaller mean absolute bias than the ML estimate. Both tables only show the cases for $\beta_{X_1R_1}^{11} < \beta_{X_2R_2}^{11}$ and for 20% of the missing percentage because the MSEs and biases are almost symmetric about the coordinate of $(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11})$. They increase as we increase the missing percentage to 30% while keeping the same patterns of the MSEs and biases as those of the missing 20%.

Table 5, where a boundary solution occurs, shows that $NON2_1^{BE}$, $NON2_3^{BE}$, $NON2_4^{BE}$ have smaller MSEs than the ML estimate (*i.e.*, $NON2_{ML}$) for all values of $\beta_{X_1R_1}^{11}$ and $\beta_{X_2R_2}^{11}$, except $(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11}) = (0.8, 0.8)$. Here, $NON2_3^{BE}$ has a smaller MSE than the ML estimate. This is true for the absolute biases. On the other hand, Table 6, where no boundary solution occurs, shows that only $NON2_3^{BE}$ is comparable to the ML estimate in the MSE although it is slightly biased. In particular, $NON2_3^{BE}$ has a smaller MSE than the ML estimate as long as $\beta_{X_1R_1}^{11} \neq 0.8$ or $\beta_{X_2R_2}^{11} \neq 0.8$ (*i.e.*, The response pattern between respondents and nonrespondents is not very different.).

Table 5
Ratios of mean MSEs and mean absolute biases of Bayesian estimates relative to the ML estimate when boundary solutions occur under a 20% missing percentage (the ratios for absolute biases are in parentheses)

	$(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
For $\{m_{j11} + m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$	(0.2, 0.2)	0.68(0.66)	0.47(0.22)	0.76(0.76)	0.65(0.48)	0.42(0.05)
	(0.2, 0.4)	0.68(0.48)	0.57(0.20)	0.77(0.68)	0.60(0.29)	0.56(0.30)
	(0.2, 0.6)	0.67(0.23)	0.73(0.66)	0.77(0.57)	0.64(0.10)	0.69(0.64)
	(0.2, 0.8)	0.77(0.26)	1.08(1.55)	0.83(0.43)	0.76(0.28)	0.95(1.34)
	(0.4, 0.4)	0.65(0.32)	0.69(0.57)	0.76(0.63)	0.61(0.17)	0.65(0.52)
	(0.4, 0.6)	0.58(0.14)	0.83(0.90)	0.71(0.56)	0.56(0.06)	0.69(0.71)
	(0.4, 0.8)	0.75(0.36)	1.46(2.07)	0.78(0.36)	0.74(0.42)	1.12(1.61)
	(0.6, 0.6)	0.66(0.22)	1.35(1.73)	0.73(0.43)	0.66(0.16)	1.01(1.29)
	(0.6, 0.8)	0.85(0.87)	2.27(3.19)	0.76(0.17)	0.83(0.81)	1.52(2.35)
	(0.8, 0.8)	1.12(1.93)	3.58(5.49)	0.83(0.24)	1.04(1.67)	2.18(3.95)
For $\{m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$	(0.2, 0.2)	0.57(0.63)	0.27(0.13)	0.69(0.74)	0.41(0.40)	0.28(0.31)
	(0.2, 0.4)	0.54(0.46)	0.37(0.34)	0.68(0.68)	0.42(0.24)	0.44(0.57)
	(0.2, 0.6)	0.51(0.19)	0.69(0.94)	0.65(0.55)	0.47(0.10)	0.69(0.88)
	(0.2, 0.8)	0.63(0.35)	1.39(2.08)	0.71(0.34)	0.62(0.47)	1.11(1.52)
	(0.4, 0.4)	0.49(0.35)	0.54(0.64)	0.65(0.64)	0.42(0.17)	0.57(0.76)
	(0.4, 0.6)	0.48(0.17)	0.98(1.24)	0.62(0.51)	0.45(0.17)	0.85(1.04)
	(0.4, 0.8)	0.62(0.44)	1.81(2.33)	0.67(0.35)	0.61(0.55)	1.35(1.81)
	(0.6, 0.6)	0.55(0.42)	1.70(1.90)	0.63(0.41)	0.54(0.40)	1.28(1.51)
	(0.6, 0.8)	0.78(0.92)	2.91(3.43)	0.69(0.14)	0.75(0.92)	1.96(2.64)
	(0.8, 0.8)	1.13(1.96)	4.63(5.72)	0.75(0.33)	1.02(1.77)	2.86(4.24)

Table 6
Ratios of mean MSEs and mean absolute biases of Bayesian estimates relative to the ML estimate when no boundary solution occurs under a 20% missing percentage (the ratios for absolute biases are in parentheses)

	$(\beta_{x_1r_1}^{11}, \beta_{x_2r_2}^{11})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
For $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$	(0.2, 0.2)	0.99(3.37)	1.05(7.00)	0.94(2.51)	0.93(4.89)	1.06(8.96)
	(0.2, 0.4)	0.98(2.57)	1.21(5.13)	0.97(1.89)	1.00(3.26)	1.24(5.56)
	(0.2, 0.6)	1.04(2.18)	1.52(3.84)	0.95(1.67)	1.06(2.38)	1.43(3.71)
	(0.2, 0.8)	1.12(2.04)	1.75(3.53)	1.00(1.48)	1.13(2.14)	1.52(3.21)
	(0.4, 0.4)	1.03(2.40)	1.49(4.66)	0.97(1.69)	1.05(2.74)	1.39(4.46)
	(0.4, 0.6)	1.20(2.17)	2.11(3.85)	1.00(1.52)	1.22(2.24)	1.78(3.42)
	(0.4, 0.8)	1.28(2.09)	2.36(3.67)	1.05(1.45)	1.26(2.09)	1.86(3.12)
	(0.6, 0.6)	1.22(2.16)	2.49(3.90)	0.96(1.48)	1.21(2.15)	1.90(3.32)
	(0.6, 0.8)	1.52(1.99)	3.19(3.39)	1.11(1.38)	1.45(1.91)	2.29(2.77)
	(0.8, 0.8)	1.66(1.96)	3.64(3.27)	1.14(1.36)	1.52(1.83)	2.43(2.59)
For $\{m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$	(0.2, 0.2)	0.88(2.59)	0.89(5.66)	0.87(2.26)	0.89(4.55)	1.21(8.69)
	(0.2, 0.4)	0.93(2.40)	1.27(4.86)	0.93(1.78)	1.00(3.08)	1.50(5.29)
	(0.2, 0.6)	1.09(2.11)	1.93(3.97)	0.98(1.40)	1.15(2.29)	1.85(3.61)
	(0.2, 0.8)	1.24(2.13)	2.36(3.90)	1.02(1.48)	1.27(2.18)	2.06(3.19)
	(0.4, 0.4)	1.03(2.18)	1.81(4.30)	0.96(1.60)	1.12(2.62)	1.85(4.39)
	(0.4, 0.6)	1.23(2.28)	2.62(4.28)	0.99(1.48)	1.29(2.42)	2.28(3.80)
	(0.4, 0.8)	1.42(2.05)	3.26(3.70)	1.07(1.42)	1.44(2.07)	2.53(3.09)
	(0.6, 0.6)	1.33(2.07)	3.22(3.95)	0.99(1.36)	1.36(2.14)	2.54(3.43)
	(0.6, 0.8)	1.65(2.09)	4.14(3.74)	1.13(1.43)	1.61(2.07)	2.98(3.13)
	(0.8, 0.8)	1.91(2.02)	4.48(3.50)	1.16(1.39)	1.66(1.93)	3.03(2.83)

Park and Brown (1994) used $NON2_2^{BE}$ to estimate expected cell counts in an incomplete one-way table under a nonignorable nonresponse mechanism. They showed by simulation studies that $NON2_2^{BE}$ has a smaller MSE than the ML estimate although it is biased more than the ML. However, larger values than 1 for $NON2_2^{BE}$ in Table 5 and Table 6 indicate that this is not true in an incomplete two-way table regardless of the boundary solution and that Bayesian methods are not always better than the ML even when a boundary solution occurs. A reason that our simulation results differ from those of Park and Brown (1994) when a boundary solution occurs is attributed to the choice of $(\beta_{x_1r_1}^{11}, \beta_{x_2r_2}^{11})$ where Park and Brown performed their simulation only for $\beta_{x_1r_1}^{11} = \beta_{x_2r_2}^{11} = 0.34$. As shown in Table 5, $NON2_2^{BE}$ is better than the ML when $\beta_{x_1r_1}^{11} \leq 0.4$ and $\beta_{x_2r_2}^{11} \leq 0.4$, whereas $NON2_2^{BE}$ is worse than the ML when the response pattern between respondents and nonrespondents is much different (i.e., $\beta_{x_1r_1}^{11} \geq 0.6$ or $\beta_{x_2r_2}^{11} \geq 0.6$).

Table 7 provides the mean of the standard deviations and the 95% coverage probabilities for $\beta_{x_1r_1}^{11}$. Here, we used the variance formula given in (9) to calculate the standard

deviations and the 95% coverage probabilities are the coverage rates for nominal 95% confidence intervals. When a boundary solution occurs, although the coverage probability of the ML estimate is closest to the 95% nominal coverage level, the ML estimate has too large a standard deviation to use in practice. Such large standard deviations are due to the boundary problem of the ML estimate. The coverage probabilities of $NON2_3^{BE}$ are the closest to the 95% nominal coverage level among the Bayesian estimates, while those of the other Bayesian estimates are generally smaller than the 95% nominal coverage level. This implies that the Bayesian estimates other than $NON2_3^{BE}$ underestimate their standard deviations.

When no boundary solution occurs (the second table in Table 7), the standard deviations of the ML estimate are much more stable, compared to those for the boundary solution case. The coverage probability decreases as $\beta_{x_1r_1}^{11}$ and $\beta_{x_2r_2}^{11}$ increase. In particular, the coverage probabilities of $NON1_1^{BE}$, $NON2_2^{BE}$, and $NON5_5^{BE}$ are seriously smaller than the 95% nominal coverage level when the response pattern between the respondents and undecided voters is much different (i.e., $\beta_{x_1r_1}^{11} \geq 0.6$ and $\beta_{x_2r_2}^{11} \geq 0.6$).

Table 7
Mean of standard deviations and 95% coverage probabilities (in parentheses) for $\beta_{X_1 R_1}^{11}$

	$(\beta_{X_1 R_1}^{11}, \beta_{X_2 R_2}^{11})$	$NON2_{ML}$	$NON2_1^{RE}$	$NON2_2^{RE}$	$NON2_3^{RE}$	$NON2_4^{RE}$	$NON2_5^{RE}$
boundary	(0.2, 0.2)	89.5(0.974)	0.082(0.978)	0.064(0.978)	0.093(0.973)	0.071(0.972)	0.060(0.957)
	(0.2, 0.4)	158.3(0.959)	0.072(0.963)	0.096(0.963)	0.079(0.958)	0.066(0.958)	0.058(0.940)
	(0.2, 0.6)	135.3(0.940)	0.065(0.941)	0.057(0.941)	0.071(0.941)	0.062(0.939)	0.056(0.922)
	(0.2, 0.8)	57.4(0.930)	0.070(0.938)	0.061(0.935)	0.076(0.928)	0.066(0.928)	0.060(0.908)
	(0.4, 0.4)	153.4(0.961)	0.079(0.920)	0.061(0.913)	0.096(0.956)	0.072(0.949)	0.060(0.911)
	(0.4, 0.6)	82.2(0.955)	0.072(0.893)	0.059(0.883)	0.086(0.951)	0.069(0.940)	0.058(0.874)
	(0.4, 0.8)	51.2(0.933)	0.071(0.862)	0.059(0.849)	0.084(0.926)	0.068(0.917)	0.059(0.846)
	(0.6, 0.6)	175.5(0.946)	0.077(0.820)	0.060(0.781)	0.101(0.943)	0.074(0.921)	0.061(0.823)
	(0.6, 0.8)	159.6(0.924)	0.071(0.728)	0.057(0.657)	0.089(0.913)	0.069(0.880)	0.058(0.737)
	(0.8, 0.8)	72.8(0.920)	0.070(0.572)	0.056(0.330)	0.093(0.900)	0.070(0.842)	0.058(0.607)
no-boundary	(0.2, 0.2)	0.068(0.949)	0.060(0.959)	0.056(0.959)	0.062(0.937)	0.058(0.935)	0.055(0.922)
	(0.2, 0.4)	0.066(0.960)	0.060(0.970)	0.056(0.970)	0.061(0.935)	0.058(0.931)	0.055(0.951)
	(0.2, 0.6)	0.064(0.940)	0.058(0.945)	0.055(0.945)	0.059(0.959)	0.057(0.919)	0.054(0.909)
	(0.2, 0.8)	0.069(0.933)	0.063(0.944)	0.059(0.941)	0.065(0.926)	0.062(0.925)	0.058(0.920)
	(0.4, 0.4)	0.074(0.910)	0.061(0.836)	0.055(0.828)	0.064(0.899)	0.059(0.884)	0.055(0.824)
	(0.4, 0.6)	0.074(0.915)	0.060(0.815)	0.055(0.806)	0.064(0.922)	0.059(0.879)	0.055(0.792)
	(0.4, 0.8)	0.073(0.891)	0.061(0.786)	0.056(0.771)	0.064(0.873)	0.060(0.852)	0.056(0.763)
	(0.6, 0.6)	0.078(0.859)	0.061(0.567)	0.055(0.470)	0.067(0.853)	0.061(0.795)	0.056(0.572)
	(0.6, 0.8)	0.076(0.843)	0.060(0.515)	0.054(0.402)	0.065(0.817)	0.060(0.767)	0.055(0.556)
	(0.8, 0.8)	0.080(0.755)	0.059(0.110)	0.053(0.017)	0.065(0.728)	0.059(0.607)	0.055(0.158)

5. Concluding remarks

We investigated the Bayesian analysis for incomplete two-way contingency tables with nonignorable non-response. In this situation, the ML estimates often fall on the boundary solution. These boundary solutions can yield $G^2 > 0$ even for a saturated model (Baker *et al.* 1992; Park and Brown 1994). This means that the G^2 may not be appropriate as a statistic for model specification. To avoid the boundary solution problem and to obtain a statistic such as a Bayes factor for model specification regardless of a boundary solution, we proposed Bayesian estimation methods using five different priors. Two of them are new and the remaining three have been previously used for analyzing an incomplete one-way table. These two new priors accommodate different response patterns between respondents and nonrespondents.

Data analysis shows that these new two priors are more reasonable in the sense that they accommodate the nonignorable nonresponse mechanism better and produce estimates close to the actual results. Moreover, with the

previous three priors, our simulation study shows that the Bayesian estimates can have larger MSEs than those of the ML estimates for a contingency table with no boundary solution and a boundary solution as well, contrary to the previous studies. However, when a boundary solution occurs, the two new priors perform better than the previous three priors and the ML estimates in the sense that they have generally smaller MSEs, smaller biases, and coverage probabilities closer to the nominal coverage level.

We have briefly discussed the weighting issues at Section 2.2. However, these issues need much more rigorous discussion than we did in that section. Our discussion can be further extended to include not only different weights but also response biases and other sources of biases and variations. These problems can be carefully developed on an extended paper at a later time.

Acknowledgements

This research was supported by a Korea University Grant (K0822301).

References

- Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition. New York: John Wiley & Sons, Inc.
- Baker, S.G., and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.
- Baker, S.G., Rosenberger, W.F. and Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11, 643-657.
- Chen, T. (1972). Mixed-up frequencies and missing data in contingency tables. Unpublished Ph.D. dissertation, University of Chicago, Dept. of Statistics.
- Chen, Q.L., and Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Technical Report*, Department of Statistics, The Ohio State University.
- Clarke, P.S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response. *Biometrical Journal*, 44, 701-717.
- Clogg, C.C., Rubin, D.B., Schenker, N. and Schultz, B. (1991). Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.
- De Heer W. (1999). International response trends of an international survey. *Journal of Official Statistics*, 15, 129-142.
- Dempster, A.P., Laird, N.M. and Rubin, D.M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- Flannelly, K.J., Flannelly, L.T. and McLeod, M.S. Jr. (2000). Reducing undecided voters and other sources of error in election surveys. *International Journal of Market Research*, 42, 231-237.
- Fenwick, I, Wiseman, F, Becker, J.F. and Heiman, J.R. (1982). Classifying undecided voters in pre-election polls. *Public Opinion Quarterly*, 46, 383-391.
- Forster, J.J., and Smith, R.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society B*, 60, 57-70.
- Gelman, A., Carlin, J.P., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. 2nd Edition. New York: Chapman and Hall/CRC.
- Groves, R.M., and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.
- Kim, T. (1995). Discriminant analysis as a prediction tool for uncommitted voters in pre-election polls. *International Journal of Public Opinion Research*, 7, 110-127.
- Lau, R.R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential elections. *Public Opinion Quarterly*, 59, 589-605.
- Lavrakas, P.J. (1993). *Telephone Survey Method: Sampling, selection, and supervision*. 2nd Edition. Newbury Park, Calif.: Sage.
- Little, J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data*. 2nd Edition. New York: John Wiley & Sons, Inc.
- Martin, E.A., Traugott, M.W. and Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *The Public Opinion Quarterly*, 69, 342-369.
- Michiels, B., and Molenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom drop-out. *Communications in Statistics: Theory and Methods*, 26, 65-94.
- Molenberghs, G., Kenward, M.G. and Goetghebuer, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Applied Statistics*, 50, 15-29.
- Monterola, C., Lim, M., Garcia, F. and Saloma, C. (2001). Feasibility of a neural network as classifier of undecided respondents in a public opinion survey. *International Journal of Public Opinion Research*, 14, 222-299.
- Myers, D.J., and O'Connor, R.E. (1983). The undecided respondents in mandatory voting settings: A Venezuelan exploration. *The Western Political Quarterly*, 36, 420-433.
- Park, T., and Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, 89, 44-52.
- Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54, 1579-1690.
- Perry, P. (1979). Certain problem in election survey methodology. *Public Opinion Quarterly*, 43, 312-325.
- Potthoff, R.F. (1994). Telephone sampling in epidemiologic research: To reap the benefits, avoid the pitfalls. *American Journal of Epidemiology*, 139, 967-978.
- Rubin, D.B., Stern, H.S. and Vehovar, V. (1995). Handling "Don't Know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.
- Smith, P.W.F., Skinner, C.J. and Clarke, P.S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intention data. *Applied Statistics*, 48, 563-577.
- Smith, T.W. (1984). Non attitudes: A review and evaluation. In *Surveying Subjective Phenomena*, (Eds. C.F. Turner and E. Martin), New York: Russell Sage Foundation, 2, 215-255.
- Stasny, E.A. (1986). Estimating gross flow using panel data with nonresponse: An example from the Canadian Labor Force survey. *Journal of the American Statistical Association*, 81, 42-47.
- Stasny, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating Gross Labor-Force flows. *Journal of Business and Economic Statistics*, 6, 207-219.

Hierarchical and empirical Bayes small domain estimation of the proportion of persons without health insurance for minority subpopulations

Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff and Van L. Parsons¹

Abstract

The paper considers small domain estimation of the proportion of persons without health insurance for different minority groups. The small domains are cross-classified by age, sex and other demographic characteristics. Both hierarchical and empirical Bayes estimation methods are used. Also, second order accurate approximations of the mean squared errors of the empirical Bayes estimators and bias-corrected estimators of these mean squared errors are provided. The general methodology is illustrated with estimates of the proportion of uninsured persons for several cross-sections of the Asian subpopulation.

Key Words: Asian; Bias-corrected; Mean squared error; Second order accurate.

1. Introduction

The main motivation behind this work was small domain estimation of the proportion of individuals without health insurance for different minority subpopulations. The small domains were constructed on the basis of age, sex, race and the region where the person belongs. The National Health Interview Survey (NHIS) data provide the individual level binary response (that is whether or not a person has health insurance) along with individual level covariates. The data can be obtained at <http://www.cdc.gov/nchs/nhis.htm>. The design of NHIS is discussed in Botman, Moore, Moriarity and Parsons (2000).

In a typical year the NHIS samples dwelling units, the collective members of each unit being referred to as a household, and members with a "strong" relationship being referred to as a family. (Structural units are more explicitly defined in Chapter 5.2 in the Census document at www.census.gov/prod/2002pubs/tp63rv.pdf). Each year the NHIS data contain about 40,000 households, of which over 98% are one-family households, and contain about 100,000 persons. For "family-type" questions, *e.g.*, on insurance coverage, all adults at home are invited to participate in the interview, but proxy adult response is also allowed. Children require an adult proxy.

The original survey for any given year contains data on more than 100,000 individuals and on over 800 variables. Of these individuals, we have information on the primary response variable, namely whether a person has health insurance or not. In addition, there is information on demographic characteristics such as age, sex, race, region, education, income status, medical condition, disability conditions (if any) and many other socio-economic factors.

For the entire US population, the direct estimates for these domains, namely the weighted sample proportions, are fairly reliable, since the sample size for each domain is reasonably large. This need not be the case though when our analysis is targeted towards specific subpopulations, such as Hispanics, Asians and similar minority sectors of the community.

For a targeted minority subpopulation, the sample size in a domain is not always very large. Hence, the direct estimates may not be very reliable, being accompanied with large standard errors and coefficients of variation. This calls for the use of small domain estimation techniques, where indirect estimates are obtained for these domains based on implicit or explicit models. These models help building a link between these domains, and thus produce typically estimates of greater precision by borrowing strength.

We employ both hierarchical Bayes (HB) and empirical Bayes (EB) methodology to obtain small domain estimates and find also the associated measures of precision. The analysis is based on a HB analogue of the generalized linear mixed model (GLMM) to obtain posterior means and posterior standard errors of the population small domain proportions. The method was proposed in Ghosh, Natarajan, Stroud and Carlin (1998). The EB approach is based on the theory of optimal estimating functions. We obtain EB estimators and the corresponding approximate mean squared error estimators by an asymptotic method analogous to that of Prasad and Rao (1990) and Ghosh and Maiti (2004). While the procedure of Ghosh and Maiti (2004) is based on area-level data, the present approach uses unit level data. Hence, by necessity, one needs some modification of the procedure proposed in Ghosh and Maiti (2004) in developing the estimators. Also, the general methodology,

1. Malay Ghosh, University of Florida; Dalho Kim, Kyungpook National University; Karabi Sinha, University of California at Los Angeles; Tapabrata Maiti, Michigan State University; Myron Katzoff, National Center for Health Statistics; Van L. Parsons, National Center for Health Statistics.

like that of Ghosh and Maiti is not restricted only to binary data. The methodology is applicable to the natural exponential family with quadratic variance functions. (Morris 1982, 1983). The development of mean squared errors of the estimates under the proposed model is somewhat simpler than that of Ghosh and Maiti (2004) for the binary case. Moreover, like Ghosh and Maiti (2004), our analysis utilizes the survey weights along with the model to derive the small domain estimates. Thus, our method, in some sense, can be regarded as design-assisted model-based estimation.

Survey weights attached to individual sampling units are usually proportional to inverses of their selection probabilities. They are often used to produce design-unbiased estimators. The classic example is the celebrated Horvitz-Thompson estimator. However, while such estimators guard against model failure, they may result in loss of efficiency if the assumed model is true. For example, in a simple Bayesian set up, if $y_i | \theta_i$ are independently distributed $N(\theta_i, 1)$, while θ_i are independently and identically distributed $N(\mu, A)$, ($i = 1, \dots, n$), then the Bayes estimator (posterior mean) of $\bar{\theta} = n^{-1} \sum_{i=1}^n \theta_i$ is $n^{-1} \sum_{i=1}^n [(1-B)y_i + B\mu] = (1-B)\bar{y} + B\mu$, where $B = (1+A)^{-1}$. This estimator has Bayes risk $n^{-1}(1-B)$ under the assumed model. On the other hand, the estimator $\sum_{i=1}^n w_i y_i$ of $\bar{\theta}$, with $\sum_{i=1}^n w_i = 1$ has Bayes risk $n^{-1}(1-B) + E[(1-B)\bar{y} + B\mu - \sum_{i=1}^n w_i y_i]^2$. If, however, the assumed model is not true, for example, θ_i are independently and identically distributed $N(\mu, A)$, ($i = 1, \dots, n$), where A depends widely from A_0 , then the Bayes risk of the estimator $(1-B)\bar{y} + B\mu$ of $\bar{\theta}$ has Bayes risk $n^{-1}(1-B_0) + (B-B_0)^2(\bar{y}-\mu)^2$, $B_0 = (1+A_0)^{-1}$, which can be quite larger than the corresponding Bayes risk of $\sum_{i=1}^n w_i y_i$ depending of course on B_0 , μ and the w_i .

The present paper produces small domain estimates of the proportion of uninsured persons for the Asian population. The estimates and measures of precision are based both on the hierarchical Bayesian model as well as the EB model. The analysis was done for all the individual years 1997-2000. For brevity, the results are reported only for the year 2000. We carried out a similar analysis for the Hispanic population also. In this case, the number of small domains was 336. Since the methodology was the same as that for the Asians, to save space, we have not included in this paper that analysis as well.

The Asian group is formally composed of the (1) Chinese, (2) Filipino, (3) Asian Indian, and (4) others such as Koreans, Vietnamese, Japanese, Hawaiian, Samoan, Guamanian *etc.* These individuals are assigned to specific domains depending on their age, race, gender and the region

they come from. There are 3 age-groups (0-17, 18-64 and 65+), 2 Genders, 4 Races and 4 Regions depending on the size of the Metropolitan Statistical Area ($< 499,999$; $500,000-999,999$; $1,000,000-2,499,999$ $> 2,500,000$). Thus, the total number of domains equals $3 \times 2 \times 4 \times 4 = 96$. When the individuals are distributed to their respective domains, it turns out that many of the domains contain only a few observations. Indeed, there are several domains with a sample of size 1, while one domain has sample size zero.

The outline of the remaining sections is as follows. Section 2 addresses the selection of covariates for the Asians. Section 3 discusses the general HB methodology needed for obtaining the small domain estimates and the associated measures of precision. Section 4 discusses the adequacy of the proposed HB model. Section 5 discusses an alternative EB methodology, finds second order correct (to be made precise later) mean squared errors (MSE's) of the proposed EB estimators, and also second order correct approximation of these MSE's. Section 6 finds the small domain estimates and the corresponding measures of precision for the Asian subpopulation in 2000 using both the HB and the EB methodology, and these estimates are compared with the direct estimates. Some concluding remarks are made in Section 7.

2. Selection of covariates

As mentioned in the introduction, the number of covariates exceeds 800. Inclusion of all of them in the initial model is impractical and unnecessary. We started with what we deemed to be a meaningful set of 6 covariates and used a fully stepwise selection process (with a significance level of 0.05) to finally come up with the best model.

The six covariates that we considered were: (1) legal marital status, (2) family size, (3) education level, (4) total earnings from the previous year, (5) total family income, and (6) full time working status.

After the stepwise procedure, our final model included, along with the intercept term, the covariates family size, education level, and total family income.

Since the SURVEYREG procedure in SAS Version 8 fits linear regression models and produces hypothesis tests and estimates for survey data, we used this procedure for our covariate selection. Logistic regression for covariate selection was not available at the time when this research was done. It may be noted though that SURVEYREG accounts for clustering and unequal weighting, and produces standard errors that correctly account for complex survey designs.

3. Hierarchical Bayesian analysis

A general one-parameter exponential family model is given by

$$f(y_{ij} | \theta_{ij}) = \exp[\xi_{ij} \{y_{ij} \theta_{ij} - \psi(\theta_{ij})\}] h(y_{ij}; \xi_{ij}), \quad (3.1)$$

$j = 1, \dots, n_i$, $i = 1, \dots, k$. Here y_{ij} is the response of the j^{th} unit in the i^{th} small domain, while ξ_{ij} , the “so-called” overdispersion parameters are assumed to be known, and are taken as 1 without loss of generality. This is because one can otherwise work with the transformed parameters. $\xi_{ij} = \xi_{ij} \theta_{ij}$. The function h is a positive function which depends on the y_{ij} , but not on the θ_{ij} . If y_{ij} is binary with success probability p_{ij} , then $\theta_{ij} = \log(p_{ij})$. In our example, y_{ij} , the response of the j^{th} individual in the i^{th} small domain, is 1 or 0 depending on whether the person does not or does have health insurance. We are interested in estimation of $\bar{\mu}_{i\cdot w} = \sum_{j=1}^{n_i} w_{ij} p_{ij}$, the domain specific weighted averages of the population proportions. In this case, the direct estimator of $\bar{\mu}_{i\cdot w}$ is $\sum_{j=1}^{n_i} w_{ij} y_{ij}$. These direct estimators are usually subject to large standard errors and coefficients of variation. The survey weights w_{ij} are assumed to be known, and are normalized so that $\sum_{j=1}^{n_i} w_{ij} = 1$ for all $i = 1, \dots, k$. It must be admitted though that often in practice, the w_{ij} are only estimates, for example taking into account post-stratification and non-response. However, the actual mechanism used to generate these weights are unavailable to secondary users of the data, and we need to assume the weights to be known. Another important example, not specifically considered in this paper, is $y_{ij} \sim \text{Poisson}(\lambda_{ij})$, so that $\theta_{ij} = \log(\lambda_{ij})$. One can use a Poisson model here based on the domain level counts of uninsured people. The difficulty lies in the fact that in the present example, we have individual level and *not* domain level covariates. Modelling the counts via domain-level covariates is not possible in this situation.

In this section, we discuss how to carry out the analysis for the general hierarchical Bayesian model when we are interested in estimating $\mu_{ij} = E(y_{ij} | \theta_{ij}) = \psi'(\theta_{ij})$. Since $\psi''(\theta_{ij}) = \text{var}(y_{ij} | \theta_{ij})$, μ_{ij} is a one-to-one function of θ_{ij} . In particular, $\mu_{ij} = p_{ij}$ in the binary case. Specific applications will be considered in Section 5.

The next stage of the model is

$$\theta_{ij} = \mathbf{x}_{ij}^T \mathbf{b} + u_i; \quad j = 1, \dots, n_i, \quad i = 1, \dots, k, \quad (3.2)$$

where \mathbf{x}_{ij} are the design vectors, or equivalently the predictor vectors, \mathbf{b} is the vector of regression parameters, and u_i are the random effects. It is assumed that u_i are iid $N(0, \sigma_u^2)$. Also, let $\mathbf{X}^T = (\mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1}, \dots, \mathbf{x}_{k1}, \dots, \mathbf{x}_{kn_k})$, and assume that \mathbf{X} is a full rank matrix.

Finally, it is assumed that \mathbf{b} and σ_u^2 are mutually independent, where \mathbf{b} has the improper uniform prior on,

R^P , and σ_u^2 has an inverse gamma distribution with parameters $c/2, d/2$. i.e., $\pi(\sigma_u^2) \propto \exp(-c/2\sigma_u^2)(\sigma_u^2)^{-d/2-1}$, $c > 0$.

Let $\mathbf{y} = (y_{11}, \dots, y_{1n_1}, \dots, y_{k1}, \dots, y_{kn_k})^T$, and $\boldsymbol{\theta} = (\theta_{11}, \dots, \theta_{1n_1}, \dots, \theta_{k1}, \dots, \theta_{kn_k})^T$. Then the joint posterior is given by

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mathbf{b}, \sigma_u^2 | \mathbf{y}) &\propto \prod_{i=1}^k \prod_{j=1}^{n_i} f(y_{ij} | \theta_{ij}) \\ &\times (\sigma_u^2)^{-k/2} \exp \left[-\frac{1}{2\sigma_u^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (\theta_{ij} - \mathbf{x}_{ij}^T \mathbf{b})^2 \right] \\ &\times (\sigma_u^2)^{-d/2-1} \exp \left(-\frac{c}{2\sigma_u^2} \right). \end{aligned} \quad (3.3)$$

This is a nonconjugate Bayesian analysis, and is not implementable analytically. Instead, we use the Markov chain Monte Carlo (MCMC) numerical integration technique. In particular, we employ the Gibbs sampler. The general MCMC technique is discussed in many places. A convenient reference is Tanner (1996, Chapter 6).

In order to implement the Gibbs sampler, we need to find the full conditionals of θ_{ij} , \mathbf{b} and σ_u^2 . The full conditionals are given by

$$\begin{aligned} \sigma_u^2 | \boldsymbol{\theta}, \mathbf{b}, \mathbf{y} &\sim \text{IG} \left(\frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (\theta_{ij} - \mathbf{x}_{ij}^T \mathbf{b})^2 + c}{2}, \frac{k + d}{2} \right); \\ \mathbf{b} | \boldsymbol{\theta}, \sigma_u^2, \mathbf{y} &\sim N((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \sigma_u^2 (\mathbf{X}^T \mathbf{X})^{-1}); \\ \theta_{ij} | \mathbf{b}, \sigma_u^2, \mathbf{y} &\sim f(y_{ij} | \theta_{ij}) \exp \left[-\frac{1}{2\sigma_u^2} (\theta_{ij} - \mathbf{x}_{ij}^T \mathbf{b})^2 \right]. \end{aligned}$$

Our data analysis is based on generating samples from the above conditionals specialized to the binary case. Generation of samples from the conditionals of σ_u^2 and \mathbf{b} is standard. This is not so for the θ_{ij} , and requires the Metropolis-Hastings algorithm. For a discussion of this algorithm, we refer once again to Tanner (1996).

If $\hat{\mu}_{ij}^{(r)}$ denotes the sampled value of μ_{ij} generated from the r^{th} draw, and the number of draws is R , then the Monte Carlo estimate of $E(\mu_{ij} | \mathbf{y})$ is $R^{-1} \sum_{r=1}^R \hat{\mu}_{ij}^{(r)}$. Similarly, the Monte-Carlo estimate of $\text{var}(\mu_{ij} | \mathbf{y})$ is $R^{-1} \sum_{r=1}^R (\hat{\mu}_{ij}^{(r)})^2 - (R^{-1} \sum_{r=1}^R \hat{\mu}_{ij}^{(r)})^2$. Finally, Monte-Carlo estimate of $\text{cov}(\mu_{ij}, \mu_{ij'}) | \mathbf{y}$ is given by $R^{-1} \sum_{r=1}^R (\hat{\mu}_{ij}^{(r)} \hat{\mu}_{ij'}^{(r)}) - (R^{-1} \sum_{r=1}^R \hat{\mu}_{ij}^{(r)}) (R^{-1} \sum_{r=1}^R \hat{\mu}_{ij'}^{(r)})$. Based on these calculations, it is now immediate to find $E[\bar{\mu}_{i\cdot w} | \mathbf{y}] = \sum_{j=1}^{n_i} w_{ij} E(\mu_{ij} | \mathbf{y})$ and $V[\bar{\mu}_{i\cdot w} | \mathbf{y}] = \sum_{j=1}^{n_i} w_{ij}^2 V(\mu_{ij} | \mathbf{y}) + \sum_{1 \leq j \neq j' \leq n_i} w_{ij} w_{ij'} \text{Cov}(\mu_{ij}, \mu_{ij'} | \mathbf{y})$. In contrast, the direct unbiased estimator of $\bar{\mu}_{i\cdot w}$ is given by $\bar{y}_{i\cdot w} = \sum_{j=1}^{n_i} w_{ij} y_{ij}$. However, as noted earlier, for many of these domains, the sample sizes are so small that these unbiased estimators are subject to large standard errors and coefficients of variation.

4. Empirical Bayes estimation

Once again, let y_{ij} denote the response of the j^{th} unit in the i^{th} small domain ($j = 1, \dots, n_i; i = 1, \dots, k$). Also, we assume the exponential family model for the y_{ij} as given in (3.1), but it is assumed in addition that the y_{ij} has a probability function or a probability density function belonging to the natural exponential family quadratic variance function (NEF-QVF) class. We may recall that $\mu_{ij} = E(y_{ij} | \theta_{ij}) = \psi'(\theta_{ij})$. With the quadratic variance function structure, $\text{Var}(y_{ij} | \theta_{ij}) = Q(\mu_{ij}) = v_0 + v_1 \mu_{ij} + v_2 \mu_{ij}^2$, where v_0, v_1 and v_2 are not simultaneously zero. Morris (1982, 1983) has characterized distributions belonging to the NEF-QVF family. The family consists of the six basic distributions, namely, (i) Bernoulli, (ii) Poisson, (iii) normal with known variance, (iv) geometric, (v) exponential, (vi) hyperbolic secant, and their convolutions. In this way, binomial, negative binomial and gamma distributions also belong to this family. For the Bernoulli distribution, $v_0 = 0, v_1 = 1$ and $v_2 = -1$. For the Poisson distribution, $v_0 = v_2 = 0$ and $v_1 = 1$. For the normal distribution with known variance σ^2 , $\xi_{ij} = \sigma^{-2}$, $v_0 = 1$ and $v_1 = v_2 = 0$. Once again we will assume without loss of generality that $\xi_{ij} = 1$.

We propose in this section EB estimators of the small domain means. To this end, we begin with the general NEF-QVF family of distributions along with a conjugate prior for the canonical parameter of the exponential model. Together they constitute an overdispersed NEF-QVF family of distributions. Specifically, we consider the conjugate prior with pdf

$$\pi(\theta_{ij}) = \exp[\lambda \{m_{ij} \theta_{ij} - \psi(\theta_{ij})\}] C(\lambda, m_{ij}) \quad (4.1)$$

for θ_{ij} , where $m_{ij} = g(x_{ij}^T \mathbf{b})$, $j = 1, \dots, n_i; i = 1, \dots, k$. Here x_{ij} is the design vector associated with the j^{th} unit in the i^{th} small domain, and g is the link function. Then (Morris 1983),

$$E(\mu_{ij}) = m_{ij}; \text{var}(\mu_{ij}) = Q(m_{ij})/(\lambda - v_2), \quad (4.2)$$

where we assume that $\lambda > \max(0, v_2)$. Since $\text{var}(\mu_{ij})$ is strictly decreasing in λ , we may interpret the latter as the precision parameter.

We first obtain the Bayes estimator of μ_{ij} . This is given by (Morris 1983)

$$E(\mu_{ij} | y_{ij}) = \frac{1}{\lambda + 1} y_{ij} + \frac{\lambda}{\lambda + 1} m_{ij}(\mathbf{b}).$$

The above can also be viewed as the best linear unbiased predictor (BLUP) of μ_{ij} . To see this, we calculate

$$E(y_{ij}) = E(\mu_{ij}) = m_{ij}; \text{cov}(y_{ij}, \mu_{ij}) = \text{var}(\mu_{ij})$$

$$= Q(m_{ij})/(\lambda - v_2); \text{var}(y_{ij}) = \frac{\lambda + 1}{\lambda - v_2} Q(m_{ij}).$$

Hence, the BLUP of μ_{ij} is given by

$$\begin{aligned} m_{ij}(\mathbf{b}) + \frac{\text{cov}(y_{ij}, \mu_{ij})}{\text{var}(y_{ij})} (y_{ij} - m_{ij}(\mathbf{b})) \\ = \frac{1}{\lambda + 1} y_{ij} + \frac{\lambda}{\lambda + 1} m_{ij}(\mathbf{b}). \end{aligned} \quad (4.3)$$

Thus the Bayes estimator of $\bar{\mu}_{iw} = \sum_{j=1}^{n_i} w_{ij} \mu_{ij}$ is given by $\sum_{j=1}^{n_i} w_{ij} E(\mu_{ij} | y_{ij})$.

In practice, however, \mathbf{b} and λ are unknown, and need to be estimated from the marginals of the y_{ij} . However, except for the normal distribution, these marginals are fairly complicated, and finding MLE's from the marginal likelihoods can become quite formidable. Instead, we find estimates based on some optimal unbiased estimating equations (Godambe and Thompson 1989) which requires only evaluation of the first four moments of these marginals. To this end, we begin with the elementary unbiased estimating functions $ig_{1ij} = y_{ij} - m_{ij}$ and $g_{2ij} = (y_{ij} - m_{ij})^2 - (\lambda + 1)/(\lambda - v_2) V(m_{ij})$. In order to construct the optimal estimating equations, let

$$\mathbf{D}_{ij}^T = \begin{bmatrix} -E\left(\frac{\partial g_{1ij}}{\partial \mathbf{b}}\right) & -E\left(\frac{\partial g_{2ij}}{\partial \mathbf{b}}\right) \\ -E\left(\frac{\partial g_{1ij}}{\partial \lambda}\right) & -E\left(\frac{\partial g_{2ij}}{\partial \lambda}\right) \end{bmatrix}.$$

Also, let

$$\Sigma_{ij} = \begin{bmatrix} \mu_{2ij} & \mu_{3ij} \\ \mu_{3ij} & \mu_{4ij} - \mu_{2ij}^2 \end{bmatrix},$$

where $\mu_{rij} = E(y_{ij} - m_{ij})^r$ is the r^{th} central moment of y_{ij} based on its marginal distribution. The optimal estimating equations are then given by $\sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{D}_{ij}^T \Sigma_{ij}^{-1} \mathbf{g}_{ij} = \mathbf{0}$, where $\mathbf{g}_{ij} = (g_{1ij} \ g_{2ij})^T$. We obtain estimates of \mathbf{b} and λ (if they exist) by solving these equations. The solutions of these equations are found by the Nelder-Meade algorithm.

Unfortunately, the above method fails for binary data. In this case, $v_2 = -1$ so that $\text{var}(y_{ij})$ does not depend on λ . Indeed, the marginal beta-binomial distributions of the y_{ij} are unidentifiable in λ . A simple way to verify this is that if $y | p \sim \text{Bin}(1, p)$, and $p \sim \text{Beta}(\lambda m, \lambda(1 - m))$, then $E(y) = E(p) = m$, and a binary distribution is completely characterized by its mean. The problem does not occur for a Binomial(n, p) distribution with $n \geq 2$ since with the same marginal for p , the mgf of the marginal distribution of the binomial y is $E[(p \exp(t) + 1 - p)^n]$ which depends on λ .

For binary y_{ij} , $\partial g_{1ij} / \partial \lambda = \partial g_{2ij} / \partial \lambda = 0$ so that the second element of the vector $\sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{D}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{g}_{ij}$ is zero. Accordingly, the proposed estimating equations approach fails to estimate λ . The basic data, to be considered in our application, is binary, and this necessitates modification of the proposed procedure.

We have thus considered the optimal estimating function (for known λ)

$$\sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - m_{ij}) / (\text{var}(y_{ij}))] \frac{\partial m_{ij}}{\partial \mathbf{b}} = \mathbf{0},$$

since $\partial g_{1ij} / \partial \mathbf{b} = -\partial m_{ij} / \partial \mathbf{b}$. It may be noted also that in this case $\text{var}(y_{ij}) = V(m_{ij}) = m_{ij}(1 - m_{ij})$. Also, with the logistic representation, $m_{ij}(\mathbf{b}) = \exp(\mathbf{x}_{ij}^T \mathbf{b}) / [1 + \exp(\mathbf{x}_{ij}^T \mathbf{b})]$, one gets $\partial m_{ij} / \partial \mathbf{b} = -m_{ij}(1 - m_{ij})\mathbf{x}_{ij}$. Thus \mathbf{b} is estimated from the estimating equations $\sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{x}_{ij} y_{ij} = \sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{x}_{ij} m_{ij}$. Denoting this estimator by $\hat{\mathbf{b}}$, the EB estimator of μ_{ij} is given by

$$\hat{\mu}_{ij}^{\text{EB}} = \frac{1}{\lambda + 1} y_{ij} + \frac{\lambda}{\lambda + 1} m_{ij}(\hat{\mathbf{b}}). \quad (4.4)$$

Accordingly, the EB estimator of $\bar{\mu}_{iw}$ is $\hat{\bar{\mu}}_{iw}^{\text{EB}} = \sum_{j=1}^{n_i} w_{ij} \hat{\mu}_{ij}^{\text{EB}}$.

The procedure described above assumes a known λ . One can find estimates for the μ_{ij} for different choices of λ . In this article, we have tried $\lambda = 0.1, 0.5$ and 1, and have compared the estimates with the corresponding HB estimates.

Next, in this section, we find the mean squared errors (MSE) and also the estimated MSE's of $\hat{\bar{\mu}}_{iw}^{\text{EB}}$ assuming known λ . We state two theorems in this section. Some notations are needed before stating these theorems. Let $\mathbf{M} = \text{Diag}(m_{11}, \dots, m_{1n_1}, \dots, m_{k1}, \dots, m_{kn_k})$ and $\boldsymbol{\Sigma}(\mathbf{b}) = \mathbf{X}^T \mathbf{M} (\mathbf{I} - \mathbf{M}) \mathbf{X} = \sum_{i=1}^k \sum_{j=1}^{n_i} m_{ij}(1 - m_{ij}) \mathbf{x}_{ij} \mathbf{x}_{ij}^T$. Also, let $n_T = \sum_{i=1}^k n_i$. It is assumed that $1 \leq n_i \leq C$ for every i , so that $n_T = O_e(k)$, where O_e denotes the exact order. The two theorems are now given below.

Theorem 1. Assume $\boldsymbol{\Sigma}(\mathbf{b}) = O_e(k)$, i.e., each element of $\boldsymbol{\Sigma}(\mathbf{b})$ is bounded below by some constant C_1 , and is bounded above by some constant C_2 , where $0 < C_1 < C_2 < \infty$. Then an approximate expression for $\text{MSE}(\hat{\bar{\mu}}_{iw}^{\text{EB}})$ correct up to $O(k^{-1})$ is given by

$$\begin{aligned} \text{MSE}(\hat{\bar{\mu}}_{iw}^{\text{EB}}) &\doteq \frac{\lambda}{(\lambda + 1)^2} \sum_{j=1}^{n_i} w_{ij}^2 m_{ij}(b)(1 - m_{ij}(b)) \\ &+ \frac{\lambda^2}{(\lambda + 1)^2} \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(b)(1 - m_{ij}(b)) \mathbf{x}_{ij} \right]^T \\ &\times \boldsymbol{\Sigma}^{-1}(\mathbf{b}) \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(b)(1 - m_{ij}(b)) \mathbf{x}_{ij} \right]. \end{aligned} \quad (4.5)$$

Theorem 2. Assume $\boldsymbol{\Sigma}(\mathbf{b}) = O_e(k)$. Then the following approximation to $\text{MSE}(\hat{\bar{\mu}}_{iw}^{\text{EB}})$ holds correct up to $O(k^{-1})$.

$$\begin{aligned} &\frac{\lambda}{(1 + \lambda)^2} \sum_{j=1}^{n_i} \left[m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}})) - (1 - 2m_{ij}(\hat{\mathbf{b}}))m_{ij}(\hat{\mathbf{b}}) \right. \\ &\quad \left. \begin{pmatrix} \text{tr}(\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{b}})\mathbf{K}_1(\hat{\mathbf{b}})) \\ \vdots \\ \text{tr}(\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{b}})\mathbf{K}_p(\hat{\mathbf{b}})) \end{pmatrix} \right. \\ &\quad \left. + m_{ij}^2(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}}))^2 \mathbf{x}_{ij}^T \boldsymbol{\Sigma}^{-1}(\hat{\mathbf{b}}) \mathbf{x}_{ij} \right] \\ &+ \frac{\lambda^2}{(\lambda + 1)^2} \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}})) \mathbf{x}_{ij} \right]^T \\ &\times \boldsymbol{\Sigma}^{-1}(\hat{\mathbf{b}}) \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}})) \mathbf{x}_{ij} \right]. \end{aligned} \quad (4.6)$$

The proofs of these theorems are deferred to the Appendix. We will apply these results in finding approximate estimates of MSE's of EB estimators in the next section. However, before that, the following point is worth noting.

If one denotes the coefficient of $\lambda/(1 + \lambda)^2$ by $B_i(\hat{\mathbf{b}})$ and the coefficient of $\lambda^2/(1 + \lambda)^2$ by $C_i(\hat{\mathbf{b}})$ in Theorem 2, then noting that $B_i(\hat{\mathbf{b}}) = O(1)$ and $C_i(\hat{\mathbf{b}}) = O(k^{-1})$, for large k , $\text{MSE}(\hat{\bar{\mu}}_{iw}^{\text{EB}})$ is maximized at $\hat{\lambda} = (B_i(\hat{\mathbf{b}}))/(B_i(\hat{\mathbf{b}}) - 2C_i(\hat{\mathbf{b}}))$ which is typically very close to 1. The resulting prior with $\hat{\lambda}$ replacing λ is the data adaptive approximate least favorable prior. In the example to be considered, this estimated λ turns out to be 1.003 which conforms the above observation.

5. Small domain estimates for Asians

We first describe how the small domains are constructed. Consider the 4-tuple (k_1, k_2, k_3, k_4) , where $k_1 = 1, 2, 3$ or 4 according as the person is Chinese, Filipino, Asian Indian or Islanders. Next $k_2 = 1$ or 2 according as the person is a male or a female. Then $k_3 = 1, 2$ or 3 according as the person belongs to the age-group 0-17, 18-64 or 65+. Finally, $k_4 = 1, 2, 3$ or 4 according as the person belongs to a Metropolitan Statistical Area (MSA) of size $\leq 499,999$, 500,000 - 999,999, 1,000,000 - 2,499,999 or $\geq 2,500,000$. A small domain is now numbered by the formula $24(k_1 - 1) + 12(k_2 - 1) + 4(k_3 - 1) + k_4$ corresponding to the 4-tuple (k_1, k_2, k_3, k_4) . For example, the small domain consisting of Filipino females belonging to the age-group

18-64 and a MSA of size 500,000 – 999,999 is numbered 42.

The basic data consist of $y_{ij} = 1$ or 0 if the j^{th} individual in the i^{th} small domain does not (does) have health insurance;

- \tilde{w}_{ij} = the sampling weight attached to the j^{th} unit in the i^{th} small domain;
- w_{ij} = $\tilde{w}_{ij} / \sum_{j=1}^{n_i} \tilde{w}_{ij}$ so that $\sum_{j=1}^{n_i} w_{ij} = 1$ for each i .
- x_{ij1} = the family size of the j^{th} unit in the i^{th} small domain;
- x_{ij2} = the education level of the j^{th} unit in the i^{th} small domain;
- x_{ij3} = total family income of the j^{th} unit in the i^{th} small domain;

Let $p_{ij} = E(y_{ij})$. For the HB analysis, we model

$$\theta_{ij} = \text{logit}(p_{ij}) = b_0 + b_1 x_{ij1} + b_2 x_{ij2} + b_3 x_{ij3} + u_i,$$

$$j = 1, \dots, n_i, i = 1, \dots, 96.$$

The direct domain estimates are given by $\hat{p}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij}$. The corresponding hierarchical Bayes estimates are given by $\hat{p}_{iw}^{\text{HB}} = \sum_{j=1}^{n_i} w_{ij} E(p_{ij} | y)$. We use MCMC as described in Section 2 to obtain these estimates. They are referred to in the table as HB. The associated posterior standard errors are

referred to as $\text{se}(\text{HB})$. Our hyperprior considers: $c = 0.2, 0.02, 0.002$; $d = 0.2, 0.02, 0.002$. The results are very insensitive to the choice of the hyperpriors, and are reported only for $c = d = 0.02$. In addition, we have EB estimators for different choices of the parameter λ . The results are reported for $\lambda = 0.1, 0.5$ and 1.

Table 1 provides the various estimates of uninsured Asian people and the associated standard errors for the different small domains for the year 2000. Domain 2 is excluded due to zero sample size. Domain 2 refers to Male Filipinos in the age group 0-17 belonging to MSA's of size 500,000 - 999,999. The measures of precision (posterior s.d.'s) associated with the HB estimates are denoted by $\text{se}(\text{HB})$ and are given by the formula $\text{se}^2(\text{HB}) = \text{var}(\sum_{j=1}^{n_i} w_{ij} p_{ij} | y)$. One of the advantages of the HB or EB estimates is that for domains with very small sample sizes, often the direct estimates of the proportion of uninsured is zero, whereas the former provide small but non-zero estimates. We chose not to collapse the direct estimates for domains with very small sample sizes. The unit level covariates were quite distinct, and there was no meaningful way to combine them. We note also that when $\lambda = 0.5$, i.e., the direct and synthetic estimates have 1: 2 weight ratio, the EB and HB estimates are very close.

Table 1
Small domain estimates of the proportions of uninsured Asians: Year 2000

Domain	n_i	Direct	'97-'99 average	HB	se (HB)	EB $\lambda = 0.5$	EB $\lambda = 1$	se (EB) $\lambda = 0.5$	se (EB) $\lambda = 1$
1	10	0.126	0.034	0.133	0.043	0.148	0.158	0.057	0.060
2	0	-	0.085	-	-	-	-	-	-
3	24	0.063	0.016	0.074	0.025	0.076	0.082	0.037	0.039
4	28	0.146	0.105	0.150	0.027	0.163	0.171	0.041	0.043
5	20	0.138	0.265	0.143	0.032	0.153	0.160	0.043	0.046
6	17	0.112	0.124	0.120	0.032	0.134	0.144	0.019	0.021
7	78	0.097	0.100	0.104	0.015	0.107	0.112	0.022	0.024
8	66	0.274	0.229	0.253	0.023	0.240	0.224	0.072	0.076
9	5	0.173	0.000	0.164	0.061	0.160	0.154	0.078	0.082
10	6	0.000	0.000	0.033	0.051	0.082	0.123	0.070	0.074
11	7	0.000	0.084	0.032	0.047	0.090	0.134	0.054	0.057
12	11	0.335	0.000	0.302	0.056	0.275	0.245	0.060	0.064
13	7	0.134	0.061	0.134	0.045	0.130	0.128	0.103	0.110
14	2	0.000	0.151	0.020	0.064	0.026	0.039	0.031	0.033
15	27	0.000	0.104	0.023	0.023	0.035	0.052	0.032	0.034
16	29	0.113	0.191	0.119	0.024	0.123	0.127	0.033	0.035
17	27	0.120	0.223	0.127	0.025	0.141	0.152	0.044	0.047
18	14	0.000	0.106	0.024	0.030	0.041	0.062	0.019	0.021
19	77	0.131	0.111	0.133	0.015	0.133	0.134	0.021	0.023
20	75	0.223	0.222	0.213	0.018	0.207	0.200	0.089	0.095
21	3	0.000	0.000	0.022	0.056	0.028	0.043	0.070	0.074
22	6	0.000	0.184	0.026	0.045	0.052	0.079	0.071	0.075
23	8	0.000	0.022	0.037	0.050	0.108	0.162	0.063	0.067
24	9	0.000	0.000	0.029	0.042	0.062	0.093	0.052	0.055
25	10	0.000	0.083	0.023	0.034	0.031	0.046	0.061	0.065
26	6	0.000	0.018	0.020	0.039	0.029	0.044	0.031	0.033
27	32	0.098	0.041	0.105	0.023	0.108	0.114	0.035	0.037
28	23	0.000	0.092	0.024	0.025	0.037	0.055	0.032	0.034
29	25	0.187	0.211	0.173	0.030	0.151	0.134	0.035	0.037
30	23	0.227	0.076	0.210	0.032	0.188	0.169	0.021	0.022

Table 1 (continued)

Small domain estimates of the proportions of uninsured Asians: Year 2000

Domain	n_i	Direct	'97-'99 average	HB	se (HB)	EB $\lambda = 0.5$	EB $\lambda = 1$	se (EB) $\lambda = 0.5$	se (EB) $\lambda = 1$
31	71	0.118	0.059	0.123	0.016	0.125	0.128	0.024	0.026
32	50	0.109	0.156	0.113	0.019	0.112	0.113	0.113	0.120
33	2	0.000	0.000	0.024	0.071	0.037	0.055	0.115	0.122
34	2	0.000	0.000	0.026	0.073	0.047	0.070	0.058	0.061
35	8	0.108	0.000	0.113	0.042	0.112	0.114	0.067	0.071
36	7	0.000	0.000	0.030	0.045	0.065	0.098	0.051	0.054
37	9	0.062	0.197	0.069	0.035	0.062	0.063	0.036	0.038
38	17	0.000	0.019	0.019	0.024	0.023	0.034	0.037	0.040
39	24	0.117	0.022	0.124	0.028	0.134	0.142	0.040	0.043
40	20	0.000	0.070	0.028	0.029	0.052	0.078	0.025	0.027
41	50	0.163	0.145	0.160	0.020	0.156	0.153	0.027	0.029
42	38	0.141	0.114	0.139	0.021	0.133	0.130	0.020	0.022
43	76	0.104	0.090	0.112	0.016	0.120	0.128	0.020	0.022
44	73	0.142	0.149	0.142	0.016	0.139	0.137	0.119	0.127
45	2	0.000	0.000	0.027	0.076	0.051	0.076	0.090	0.095
46	3	0.000	0.052	0.021	0.056	0.023	0.035	0.052	0.055
47	10	0.000	0.072	0.024	0.034	0.044	0.066	0.068	0.072
48	7	0.000	0.172	0.029	0.045	0.068	0.102	0.051	0.054
49	10	0.087	0.364	0.095	0.037	0.099	0.105	0.078	0.083
50	5	0.000	0.000	0.027	0.050	0.053	0.080	0.032	0.034
51	23	0.038	0.092	0.053	0.023	0.056	0.066	0.037	0.039
52	21	0.243	0.195	0.223	0.037	0.198	0.176	0.030	0.032
53	31	0.114	0.184	0.120	0.022	0.121	0.124	0.040	0.042
54	18	0.202	0.169	0.195	0.031	0.188	0.182	0.019	0.020
55	74	0.094	0.115	0.102	0.015	0.102	0.106	0.019	0.020
56	83	0.204	0.296	0.192	0.017	0.178	0.165	0.133	0.141
57	2	0.000	0.124	0.029	0.082	0.062	0.092	0.146	0.154
58	1	0.000	0.000	0.019	0.087	0.023	0.035	0.000	0.000
59	2	0.000	0.196	0.020	0.063	0.021	0.032	0.103	0.194
60	8	0.112	0.116	0.120	0.044	0.132	0.143	0.059	0.063
61	16	0.202	0.140	0.187	0.036	0.169	0.152	0.040	0.043
62	3	0.301	0.163	0.276	0.086	0.252	0.227	0.100	0.107
63	33	0.055	0.093	0.069	0.020	0.073	0.082	0.028	0.030
64	28	0.105	0.275	0.112	0.024	0.115	0.120	0.032	0.034
65	33	0.126	0.133	0.129	0.021	0.126	0.126	0.029	0.031
66	13	0.393	0.290	0.350	0.054	0.323	0.288	0.048	0.051
67	70	0.080	0.136	0.089	0.015	0.088	0.093	0.019	0.021
68	75	0.179	0.233	0.171	0.017	0.159	0.149	0.019	0.021
69	1	0.000	0.000	0.851	0.248	0.705	0.558	0.163	0.173
70	2	0.361	0.000	0.331	0.098	0.299	0.268	0.119	0.126
71	4	0.000	0.091	0.023	0.050	0.032	0.048	0.077	0.082
72	2	0.000	0.182	0.045	0.101	0.157	0.236	0.155	0.165
73	45	0.271	0.144	0.256	0.026	0.256	0.249	0.028	0.030
74	10	0.000	0.044	0.024	0.034	0.034	0.051	0.051	0.055
75	83	0.149	0.097	0.150	0.016	0.160	0.166	0.020	0.021
76	59	0.113	0.205	0.120	0.018	0.128	0.136	0.023	0.024
77	68	0.338	0.224	0.313	0.025	0.302	0.284	0.023	0.024
78	39	0.098	0.138	0.103	0.020	0.102	0.104	0.026	0.028
79	122	0.110	0.163	0.117	0.013	0.125	0.133	0.016	0.017
80	125	0.308	0.314	0.281	0.020	0.262	0.239	0.016	0.017
81	7	0.000	0.000	0.029	0.043	0.066	0.099	0.065	0.069
82	12	0.000	0.045	0.025	0.032	0.047	0.070	0.048	0.051
83	13	0.049	0.017	0.068	0.035	0.088	0.108	0.050	0.053
84	4	0.000	0.061	0.028	0.056	0.060	0.091	0.088	0.093
85	32	0.189	0.113	0.193	0.027	0.217	0.231	0.035	0.037
86	10	0.136	0.056	0.137	0.036	0.127	0.123	0.051	0.054
87	52	0.192	0.098	0.185	0.021	0.184	0.180	0.024	0.026
88	65	0.153	0.120	0.155	0.018	0.162	0.166	0.022	0.024
89	71	0.285	0.210	0.265	0.022	0.256	0.242	0.022	0.023
90	57	0.086	0.146	0.095	0.017	0.102	0.110	0.022	0.024
91	153	0.149	0.167	0.150	0.011	0.156	0.160	0.014	0.015
92	138	0.308	0.285	0.283	0.020	0.266	0.244	0.015	0.017
93	10	0.000	0.000	0.030	0.041	0.073	0.110	0.059	0.063
94	16	0.067	0.015	0.081	0.029	0.090	0.101	0.042	0.044
95	18	0.108	0.018	0.123	0.032	0.145	0.163	0.046	0.049
96	14	0.111	0.087	0.125	0.039	0.160	0.185	0.050	0.053

The HB estimates of the proportion of uninsured for Asians vary in the 2%-35% range for the different small domains excluding domain 69. Admittedly, the EB and HB estimates for domain 69 are very adversely affected due to small sample size. We also report the standard errors associated with the HB estimates, and estimated approximate root mean squares accompanying the EB estimates. The proposed approach largely overcomes the valid criticism that naive EB estimates of standard errors (which ignore the $O(k^{-1})$ term) are typically underestimates. We have also provided a column giving the 3-year average of the direct estimates in 1997-1999. This is primarily to examine whether domains with zero direct estimates in 2000 also possess the same feature in other years, and also for comparison of EB and HB estimates with these estimates rather than the direct estimates. It turns out that with very few exceptions, the 1997-1999 average do not conform very much to the direct estimates. However, domain 69 still has zero direct estimate.

Table 2 provides the summary table for the proportion of uninsured for the three age groups 0-17, 18-64 and 65+ individually for Chinese (Asian 1), Filipino (Asian 2), Asian Indian (Asian 3) and other Asians (Asian 4). It turns out that at this higher level of aggregation, both the EB and HB small domain estimates are fairly close to the corresponding direct estimates except possibly for the age-group 65+. This seems to be quite satisfactory, since at this level of aggregation, the direct estimates often serve as benchmarks for comparison purpose.

Table 2
Proportions without health insurance coverage by age group and Asian group in 2000

	Direct	HB	EB ($\lambda = 0.5$)	EB ($\lambda = 1$)
0-17 years				
Total	0.120	0.126	0.131	0.137
Asian 1	0.087	0.097	0.105	0.114
Asian 2	0.046	0.062	0.071	0.083
Asian 3	0.113	0.117	0.114	0.114
Asian 4	0.165	0.165	0.171	0.175
18-64 years				
Total	0.177	0.172	0.168	0.164
Asian 1	0.162	0.160	0.160	0.159
Asian 2	0.137	0.137	0.135	0.134
Asian 3	0.150	0.147	0.141	0.137
Asian 4	0.219	0.208	0.203	0.195
65+ years				
Total	0.063	0.080	0.103	0.123
Asian 1	0.083	0.097	0.123	0.143
Asian 2	0.021	0.043	0.064	0.085
Asian 3	0.119	0.126	0.136	0.145
Asian 4	0.055	0.075	0.100	0.123

6. Model diagnostics and implementation of the hierarchical Bayesian model

We followed Gelman and Rubin (1992) for the implementation and convergence diagnostics of the Gibbs sampler. In particular we took 5 chains each of size 1,000 with an initial burning period of 1,000 iterations. We checked the potential scale reduction factors for convergence and these appeared to be very close to unity ($= 1$ at convergence) for each one of the parameters. A number of other diagnostics criteria are available in the literature, and are implemented via the software CODA. A partial output is provided in the Figure 1. The left side shows the overlap of the 5 parallel chains, and the right side shows the posterior inference for each parameter and the deviance ($-2 \log$ likelihood). For details regarding the description of the software that we used, we refer to Appendix C of Gelman, Carlin, Stern and Rubin (2004).

A Bayesian way to check the fit of a model to data is to draw simulated values from the posterior predictive distribution of replicated data and compare these samples to observed data. A wide departure between the generated and the observed data indicates lack of fit of the model. Following Gelman *et al.* (2004), we calculated the Bayesian p -values for checking the goodness-of-fit of the proposed Bayesian models. The general rationale behind such calculations is as follows. Let y denote the vector of observed data, ξ the vector of unknown parameters, $f(y|\xi)$ the density of y given ξ and $\Pi(\xi|y)$, the posterior density of ξ given y . Suppose one has drawn samples $\xi^{(1)}, \dots, \xi^{(R)}$ from this posterior distribution using MCMC simulation. Simulate now R hypothetical replicates of the data, say $y^{(1)}, \dots, y^{(R)}$, where $y^{(l)}, (l = 1, \dots, R)$ is drawn from the conditional distribution of y given the simulated $\xi^{(l)}$. If the model is reasonably accurate, these hypothetical replicates should be similar to the observed data y . This is formally done by first choosing a divergence variable, say $d(y, \xi)$ which will have an extreme value if the data y are in complete disagreement with the given model. Then a p -value is estimated by the proportion of cases in which the simulated divergence variable exceeds the realized value of the same. Thus the estimated p -value (usually referred to as the posterior predictive p -value) is equal to $R^{-1} \sum_{l=1}^R I_{[d(y^{(l)}, \xi^{(l)}) \geq d(y, \xi^{(l)})]}$, where I is the usual indicator function. One way of checking the goodness of fit of the model is by a scatter plot of realized values $d(y, \xi^{(l)})$ against the predictive values $d(y^{(l)}, \xi^{(l)})$ on the same scale. A good fit is indicated by about half the points in the scatter plot falling above the 45° line, and half falling below. In other words, for large samples, the estimated p -value will not be far away from one half. Of course, one may also carry out a graphical analysis by using different

plots for different subgroups, thereby allowing visualization of possible local model failure which may otherwise be obscured in the aggregate plot.

There are several possible choices of the divergence variable d . We considered a particular one in the present case. Noting that $E(Y_{ij} | p_{ij}) = p_{ij} = \exp(\theta_{ij}) / (1 + \exp(\theta_{ij}))$, one can consider the squared standardized residuals $((y_{ij} - p_{ij}^{(l)})^2) / (p_{ij}^{(l)}(1 - p_{ij}^{(l)}))$, where $p_{ij}^{(l)} = \exp(\theta_{ij}^{(l)}) / (1 + \exp(\theta_{ij}^{(l)}))$ are the generated values of the p_{ij} from the l^{th} iteration. Then the divergence variable d is

$$d(y, p^{(l)}) = \sum_{i=1}^{95} \sum_{j=1}^{n_i} \frac{(y_{ij} - p_{ij}^{(l)})^2}{p_{ij}^{(l)}(1 - p_{ij}^{(l)})}$$

$$d(y^{(l)}, p^{(l)}) = \sum_{i=1}^{95} \sum_{j=1}^{n_i} \frac{(y_{ij}^{(l)} - p_{ij}^{(l)})^2}{p_{ij}^{(l)}(1 - p_{ij}^{(l)})}$$

Clearly, there are other possible choices of d . Gelfand and Ghosh (1998) proposed a number of divergence measures, and studied their properties.

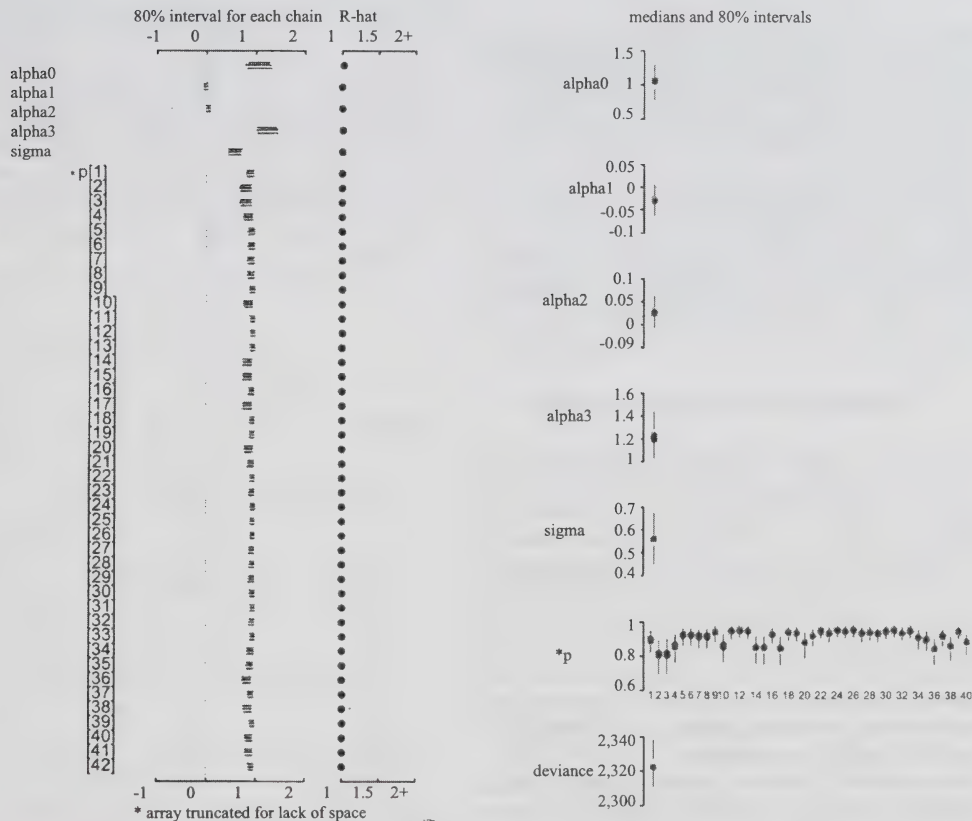


Figure 1 Bugs model at "asian_model.bug", 5 chains, each with 1,000 iterations

For the hierarchical Bayesian logistic regression model, the estimated p -value is 0.4216 for $(c, d) = (0.02, 0.02)$. The other choices of (c, d) produce similar values. The p -value bigger than 0.3 is usually treated as a good fit. Thus the proposed HB procedure seems to work well in this situation.

We have also calculated the $p_D = \text{var}(\text{deviance})/2$ and the deviance information criterion DIC or the estimated predictive deviance. The p_D can be thought of as the number of 'unconstrained parameters in the model, where a parameter is counted as 1 if it is a part of the original model (data distribution) and is 0 if it is associated with any prior distribution. The DIC is estimated as

$$\text{DIC} = 2\hat{D}(y, \theta^{(i)}) - \hat{D}(y, \hat{\theta})$$

where $\hat{D}(y, \hat{\theta})$ is the deviance calculated at the estimated parameters and $\hat{D}(y, \theta^{(i)})$ is the estimated deviance using posterior simulation. For details, see Gelman *et al.* (2004).

For our HB analysis $p_D = 56.75$ and $\text{DIC} = 2,414.41$. Usually the p_D and DIC are used as criteria of model fitting and to select the model with best predictive power. Thus, we fit also the simple logistic regression model (current model without any random effects) which means that there is no data pooling, and the estimated p_D and DIC are 22.60 and 2,379.55 respectively. The corresponding p -value is 0.3848. Thus the proposed model seems to fit the data reasonably well

7. Summary, future work and discussion

Estimating the proportion of uninsured people, especially among the minorities, is definitely a problem of great importance, and is likely to affect the policy making of Federal and State agencies. We have just started addressing this very important issue, and have provided both empirical and hierarchical Bayesian small domain estimates for the Asian subpopulation cross-classified by age, sex and other demographic characteristics. We have also discussed the adequacy of our model fit via posterior predictive p -value. Much work remains to be done however. In particular, we want to extend the present findings to the analysis of bivariate and multivariate binary data.

As pointed out by a reviewer, the present analysis ignores household clustering in the likelihood, since the original survey was a household survey, and very definitely, insurance coverage is correlated within households. However, we have assumed only a conditionally independent hierarchical model given the covariates and the random effects. Once, we have assigned distributions to the random effects, and subsequently distributions to the regression coefficients and the variance components, dependence is

built automatically in the final model, both at the unit and domain levels. Moreover, as mentioned earlier, adequacy of the hierarchical Bayesian model has been tested through posterior predictive p -values.

As a final comment, the research presented here is for illustrative purposes only. Implementation of this method for policy related matters would require further considerations of the methods and adherence to institutional standards for official policy release.

Acknowledgements

We thank the AE and two reviewers for their constructive comments on earlier drafts of the paper. The first author's research was partially funded by NSF Grants SES-0317589 and SES-0631426. The fourth author's research was partially supported by an NSF Grant SES-0318184. The research was also partially supported by NCHS/CDC under Project Number 282286285 entitled "Model-Based Subdomain Estimates". The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Appendix

Proof of Theorem 1.

$$\begin{aligned} \text{MSE}(\hat{\mu}_{iw}^{\text{EB}}) &= E(\hat{\mu}_{iw}^{\text{EB}} - \hat{\mu}_{iw})^2 = E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - p_{ij})\right)^2 \\ &= E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B + \hat{p}_{ij}^B - p_{ij})\right)^2 \\ &= E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B)\right)^2 + E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^B - p_{ij})\right)^2 \\ &\quad + 2E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B)\right)\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^B - p_{ij})\right)\right]. \end{aligned}$$

Noting that $E(p_{ij} | y) = \hat{p}_{ij}^B$,

$$\begin{aligned} &E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B)\right)\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^B - p_{ij})\right)\right] \\ &= E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B)\right) \times E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^B - p_{ij}) | y\right)\right] = 0. \end{aligned}$$

Hence,

$$\begin{aligned} \text{MSE}(\hat{\mu}_{iw}^{\text{EB}}) &= E \left(\sum_{j=1}^n w_{ij} (\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B) \right)^2 \\ &\quad + E \left(\sum_{j=1}^n w_{ij} (\hat{p}_{ij}^B - p_{ij}) \right)^2. \end{aligned} \quad (\text{A.1})$$

But

$$E \left(\sum_{j=1}^n w_{ij} (\hat{p}_{ij}^B - p_{ij}) \right)^2 = \sum_{j=1}^n w_{ij}^2 E (\hat{p}_{ij}^B - p_{ij})^2.$$

Next we calculate

$$\begin{aligned} &E (\hat{p}_{ij}^B - p_{ij})^2 \\ &= E \left[\frac{1}{\lambda+1} y_{ij} + \frac{\lambda}{\lambda+1} m_{ij}(\mathbf{b}) - p_{ij} \right]^2 \\ &= E \left[\frac{1}{\lambda+1} (y_{ij} - p_{ij}) + \frac{\lambda}{\lambda+1} (m_{ij}(\mathbf{b}) - p_{ij}) \right]^2 \\ &= \frac{1}{(\lambda+1)^2} E (y_{ij} - p_{ij})^2 + \frac{\lambda^2}{(\lambda+1)^2} E (m_{ij}(\mathbf{b}) - p_{ij})^2 \\ &\quad + \frac{2\lambda}{(\lambda+1)^2} E (y_{ij} - p_{ij}) (m_{ij}(\mathbf{b}) - p_{ij}) \\ &= \frac{1}{(\lambda+1)^2} E (p_{ij}(1-p_{ij})) + \frac{\lambda^2}{(\lambda+1)^2} V(p_{ij}) + 0 \\ &= \frac{1}{(\lambda+1)^2} \left(\frac{\lambda}{\lambda+1} m_{ij}(\mathbf{b})(1-m_{ij}(\mathbf{b})) \right) \\ &\quad + \frac{\lambda^2}{(\lambda+1)^2} \left(\frac{m_{ij}(\mathbf{b})(1-m_{ij}(\mathbf{b}))}{\lambda+1} \right) \\ &= \frac{\lambda m_{ij}(\mathbf{b})(1-m_{ij}(\mathbf{b}))}{(\lambda+1)^2}, \end{aligned}$$

so that

$$\begin{aligned} &E \left[\sum_{j=1}^n w_{ij} (\hat{p}_{ij}^B - p_{ij}) \right]^2 \\ &= \frac{\lambda}{(\lambda+1)^2} \sum_{j=1}^n w_{ij}^2 m_{ij}(\mathbf{b})(1-m_{ij}(\mathbf{b})). \end{aligned} \quad (\text{A.2})$$

Finally, we calculate,

$$\begin{aligned} &E \left[\sum_{j=1}^n w_{ij} (\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^B) \right]^2 \\ &= \frac{\lambda^2}{(\lambda+1)^2} E \left[\sum_{j=1}^n w_{ij} (m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})) \right]^2 \\ &= \frac{\lambda^2}{(\lambda+1)^2} E \left[\sum_{j=1}^n w_{ij}^2 (m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b}))^2 \right. \\ &\quad \left. + \sum_{1 \leq j \neq k \leq n} w_{ij} w_{ik} (m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})) (m_{ik}(\hat{\mathbf{b}}) - m_{ik}(\mathbf{b})) \right]. \end{aligned} \quad (\text{A.3})$$

By two-step Taylor expansion,

$$\begin{aligned} m_{ij}(\hat{\mathbf{b}}) &\doteq m_{ij}(\mathbf{b}) + \left(\frac{\partial m_{ij}(\mathbf{b})}{\partial \mathbf{b}} \right)^T (\hat{\mathbf{b}} - \mathbf{b}) \\ &\quad + \frac{1}{2} (\hat{\mathbf{b}} - \mathbf{b})^T \frac{\partial^2 m_{ij}(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}^T} (\hat{\mathbf{b}} - \mathbf{b}). \end{aligned}$$

Noting that $(\partial^2 m_{ij}(\mathbf{b})) / (\partial \mathbf{b} \partial \mathbf{b}^T) = (1 - 2m_{ij}(\mathbf{b})) m_{ij}(\mathbf{b}) (1 - m_{ij}(\mathbf{b})) \mathbf{x}_{ij} \mathbf{x}_{ij}^T$, it follows that

$$\begin{aligned} &E [m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})]^2 \\ &\doteq E \left[m_{ij}(\mathbf{b}) (1 - m_{ij}(\mathbf{b})) \mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b}) + \frac{1}{2} (\hat{\mathbf{b}} - \mathbf{b})^T \right. \\ &\quad \left. m_{ij}(\mathbf{b}) (1 - m_{ij}(\mathbf{b})) (1 - 2m_{ij}(\mathbf{b})) \mathbf{x}_{ij} \mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b}) \right]^2 \\ &= m_{ij}^2(\mathbf{b}) (1 - m_{ij}(\mathbf{b}))^2 E \left[\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b}) + \frac{1}{2} (1 - 2m_{ij}(\mathbf{b})) \right. \\ &\quad \left. (\hat{\mathbf{b}} - \mathbf{b})^T \mathbf{x}_{ij} \mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b}) \right]^2. \end{aligned} \quad (\text{A.4})$$

The first neglected term is $O_p(\|\hat{\mathbf{b}} - \mathbf{b}\|^3)$. From Sarkar and Ghosh (1998), $\hat{\mathbf{b}} - \mathbf{b}$ is asymptotically $N(0, \Sigma^{-1}(\mathbf{b}))$, where $\Sigma(\mathbf{b})$ is defined before Theorem 1. With the assumption that $\Sigma(\mathbf{b}) = O_p(k)$, it follows that $\Sigma^{-1}(\mathbf{b}) = O_p(k^{-1})$. Thus, $\|\hat{\mathbf{b}} - \mathbf{b}\| = O_p(k^{-1/2})$. Hence, the first neglected term is $O_p(k^{-3/2})$. Next, we observe that

$$\begin{aligned} &E [\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})]^2 = E [(\hat{\mathbf{b}} - \mathbf{b})^T \mathbf{x}_{ij} \mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})] \\ &= \text{tr} [\mathbf{x}_{ij} \mathbf{x}_{ij}^T E (\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})^T]. \end{aligned} \quad (\text{A.5})$$

In order to find $E[(\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})^T]$, we proceed as follows: Let $\mathbf{T}(\mathbf{b}) = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - m_{ij}(\mathbf{b}))\mathbf{x}_{ij}$ so that $\mathbf{T}(\hat{\mathbf{b}}) = \mathbf{0}$. By one-step Taylor expansion, $\mathbf{0} = \mathbf{T}(\hat{\mathbf{b}}) = \mathbf{T}(\mathbf{b}) + [\nabla \mathbf{T}(\mathbf{b})]^T (\hat{\mathbf{b}} - \mathbf{b}) + O_p(n_T^{-1})$, where

$$\begin{aligned} \nabla \mathbf{T}(\mathbf{b}) &= -\sum_{i=1}^k \sum_{j=1}^{n_i} \left(\frac{\partial m_{ij}(\mathbf{b})}{\partial \mathbf{b}} \right) \mathbf{x}_{ij}^T \\ &= -\sum_{i=1}^k \sum_{j=1}^{n_i} m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))\mathbf{x}_{ij}\mathbf{x}_{ij}^T \\ &= -\mathbf{X}^T \mathbf{M}(\mathbf{I} - \mathbf{M})\mathbf{X} \\ &= -\Sigma(\mathbf{b}). \end{aligned} \quad (\text{A.6})$$

Thus, $\hat{\mathbf{b}} - \mathbf{b} = \Sigma^{-1} \mathbf{T}(\mathbf{b}) + O_p(n_T^{-1})$. Since $V(y_{ij}) = m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))$, $V(\mathbf{T}(\mathbf{b})) = \Sigma(\mathbf{b})$. Hence $E[(\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})^T] = \Sigma^{-1}(\mathbf{b}) + O(n_T^{-3/2})$. Also, in (8.5), we have $E(\mathbf{x}_{ij}^T)(\hat{\mathbf{b}} - \mathbf{b})^2 = \text{tr}(\mathbf{x}_{ij}\mathbf{x}_{ij}^T \Sigma^{-1}(\mathbf{b})) + O_p(n_T^{-1})$. Accordingly, by (A.4) and (A.5), we have the approximation

$$E[m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})]^2 = m_{ij}^2(\mathbf{b})(1 - m_{ij}(\mathbf{b}))^2 \mathbf{x}_{ij}^T \Sigma^{-1}(\mathbf{b})\mathbf{x}_{ij} \quad (\text{A.7})$$

which is correct up to $O(n_T^{-1})$ by our assumption. Note that the neglected term

$$\begin{aligned} E[\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})(1 - 2m_{ij}(\mathbf{b}))(\hat{\mathbf{b}} - \mathbf{b})^T \mathbf{x}_{ij} \mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})] \\ = O(n_T^{-3/2}) \end{aligned}$$

since

$$\begin{aligned} E[\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})(1 - 2m_{ij}(\mathbf{b}))(\hat{\mathbf{b}} - \mathbf{b})^T \mathbf{x}_{ij} \mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})] \\ = (1 - 2m_{ij}(\mathbf{b}))E[\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})^T \mathbf{x}_{ij}] \\ = (1 - 2m_{ij}(\mathbf{b}))E[\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})]^3 \\ = O(n_T^{-3/2}). \end{aligned}$$

Similarly, note that $E[\mathbf{x}_{ij}^T (\hat{\mathbf{b}} - \mathbf{b})]^4 = O(n_T^{-2})$ and

$$\begin{aligned} E[(m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b}))(m_{ij'}(\hat{\mathbf{b}}) - m_{ij'}(\mathbf{b}))] \\ = m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))m_{ij'}(\mathbf{b})(1 - m_{ij'}(\mathbf{b}))\mathbf{x}_{ij}^T \Sigma^{-1}(\mathbf{b})\mathbf{x}_{ij'} \\ + O(n_T^{-3/2}). \end{aligned} \quad (\text{A.8})$$

This leads to

$$\begin{aligned} E \left[\sum_{j=1}^{n_i} w_{ij} (\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^{\text{B}})^2 \right] \\ = \frac{\lambda^2}{(\lambda + 1)^2} \left[\sum_{j=1}^{n_i} w_{ij}^2 m_{ij}^2(\mathbf{b})(1 - m_{ij}(\mathbf{b}))^2 \mathbf{x}_{ij}^T \Sigma^{-1}(\mathbf{b})\mathbf{x}_{ij} \right. \\ \left. + \sum_{1 \leq j \neq j' \leq n_i} w_{ij} w_{ij'} m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))m_{ij'}(\mathbf{b}) \right. \\ \left. (1 - m_{ij'}(\mathbf{b}))\mathbf{x}_{ij}^T \Sigma^{-1}(\mathbf{b})\mathbf{x}_{ij'} \right] + O(n_T^{-3/2}) \\ = \frac{\lambda^2}{(\lambda + 1)^2} \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))\mathbf{x}_{ij} \right]^T \\ \times \Sigma^{-1}(\mathbf{b}) \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))\mathbf{x}_{ij} \right] + O(n_T^{-3/2}). \end{aligned} \quad (\text{A.9})$$

Since $\Sigma^{-1}(\mathbf{b}) = O(k^{-1})$, and $n_T = O_e(k)$ by our assumption, the theorem follows from (A.2) and (A.9).

Proof of Theorem 2. We first note that $\hat{\mathbf{b}} = \mathbf{b} + O_p(n_T^{-1/2}) = \mathbf{b} + O_p(K^{-1})$ and $\Sigma^{-1}(\mathbf{b}) = O(k^{-1})$. Hence, the second term in the right hand side of (8.7) is approximated by

$$\begin{aligned} c \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}}))\mathbf{x}_{ij} \right]^T \\ \Sigma^{-1}(\hat{\mathbf{b}}) \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}}))\mathbf{x}_{ij} \right] \end{aligned} \quad (\text{A.10})$$

($c = \lambda^2 / (1 + \lambda)^2$) which is correct up to $O(k^{-1})$.

However, if we estimate $m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b}))$ simply by $m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}}))$, we will be ignoring the $O(k^{-1})$ term. Thus, we need a careful approximation of the bias $E(\hat{\mathbf{b}} - \mathbf{b})$ to achieve the desired approximation. To this end, we follow Cox and Snell (1968).

We begin with the identity

$$\begin{aligned} E[m_{ij}(\hat{\mathbf{b}})(1 - m_{ij}(\hat{\mathbf{b}}))] \\ = E[(m_{ij}(\mathbf{b}) + m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b}))(1 - m_{ij}(\mathbf{b}) + m_{ij}(\mathbf{b}) - m_{ij}(\hat{\mathbf{b}}))] \\ = m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b})) + (1 - 2m_{ij}(\mathbf{b}))E[m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})] \\ - E[m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})]^2. \end{aligned}$$

Now, again by a two-step Taylor expansion,

$$E[m_{ij}(\hat{\mathbf{b}}) - m_{ij}(\mathbf{b})] = \left[\frac{\partial m_{ij}(\mathbf{b})}{\partial \mathbf{b}} \right]^T E(\hat{\mathbf{b}} - \mathbf{b}) + \frac{1}{2} E \left[(\hat{\mathbf{b}} - \mathbf{b})^T \frac{\partial^2 m_{ij}(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}^T} (\hat{\mathbf{b}} - \mathbf{b}) \right] + O(n_T^{-3/2}).$$

In order to find $E(\hat{\mathbf{b}} - \mathbf{b})$, we proceed as follows. We begin with the second order Taylor expansion

$$0 = T_r(\hat{\mathbf{b}}) = T_r(\mathbf{b}) + \sum_{s=1}^p (\hat{b}_s - b_s) \frac{\partial T_r(\mathbf{b})}{\partial b_s} + \frac{1}{2} \sum_{s=1}^p \sum_{t=1}^p (\hat{b}_s - b_s)(\hat{b}_t - b_t) \frac{\partial^2 T_r(\mathbf{b})}{\partial b_s \partial b_t} + O(n_T^{-3/2}).$$

Taking expectations and following Cox and Snell (1968),

$$\begin{aligned} 0 &= E(T_r(\hat{\mathbf{b}})) \\ &= \sum_{s=1}^p \left[E(\hat{b}_s - b_s) E \left(\frac{\partial T_r(\mathbf{b})}{\partial b_s} \right) + \text{Cov} \left(\hat{b}_s - b_s, \frac{\partial T_r(\mathbf{b})}{\partial b_s} \right) \right] \\ &+ \frac{1}{2} \sum_{s=1}^p \sum_{t=1}^p (E(\hat{b}_s - b_s)(\hat{b}_t - b_t)) \left(\frac{\partial^2 T_r(\mathbf{b})}{\partial b_s \partial b_t} \right) \\ &+ \frac{1}{2} \sum_{s=1}^p \sum_{t=1}^p \text{Cov} \left[(\hat{b}_s - b_s)(\hat{b}_t - b_t), \left(\frac{\partial^2 T_r(\mathbf{b})}{\partial b_s \partial b_t} \right) \right] \\ &+ O(n_T^{-3/2}) = - \sum_{s=1}^p E(\hat{b}_s - b_s) \sigma_{rs} \\ &+ \sum_{s=1}^p \sum_{u=1}^p \text{Cov} \left[\sigma^{su}(\mathbf{b}) T_u(\mathbf{b}), \frac{\partial T_r(\mathbf{b})}{\partial b_s} \right] \\ &+ \frac{1}{2} \sum_{s=1}^p \sum_{t=1}^p \sigma^{st} E \left[\frac{\partial^2 T_r(\mathbf{b})}{\partial b_s \partial b_t} \right] + O(n_T^{-3/2}). \end{aligned} \quad (\text{A.11})$$

Note $\text{Cov}[\sigma^{su}(\mathbf{b}) T_u(\mathbf{b}), \partial T_r(\mathbf{b}) / \partial b_s] = 0$ since $\partial T_r(\mathbf{b}) / \partial b_s$ is a constant independent of the y_{ij} .

Similarly,

$$\text{Cov} \left[(\hat{b}_s - b_s)(\hat{b}_t - b_t), \left(\frac{\partial^2 T_r(\mathbf{b})}{\partial b_s \partial b_t} \right) \right] = 0.$$

Also, let

$$\begin{aligned} K_{rst} &= E \left[\frac{\partial^2 T_r(\mathbf{b})}{\partial b_s \partial b_t} \right] \\ &= \frac{\partial}{\partial b_t} \sum_{i=1}^k \sum_{j=1}^{n_i} -m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b})) x_{ijr} x_{ijs} \\ &= - \sum_{i=1}^k \sum_{j=1}^{n_i} (1 - 2m_{ij}(\mathbf{b})) m_{ij}(\mathbf{b})(1 - m_{ij}(\mathbf{b})) x_{ijr} x_{ijs} x_{ijt}. \end{aligned} \quad (\text{A.12})$$

Thus, one has

$$\sum_{s=1}^k \sigma_{rs} E(\hat{b}_s - b_s) = \sum_{s=1}^k \sum_{t=1}^p \sigma^{st} K_{rst}, \quad r = 1, \dots, p.$$

In matrix notations, one gets

$$\Sigma E(\hat{\mathbf{b}} - \mathbf{b}) = \frac{1}{2} \begin{pmatrix} \text{tr}(\Sigma^{-1} \mathbf{K}_1) \\ \vdots \\ \text{tr}(\Sigma^{-1} \mathbf{K}_p) \end{pmatrix}$$

where $\mathbf{K}_r = ((K_{rst}))$.

Hence,

$$E(\hat{\mathbf{b}} - \mathbf{b}) = \frac{1}{2} \Sigma^{-1} \begin{pmatrix} \text{tr}(\Sigma^{-1} \mathbf{K}_1) \\ \vdots \\ \text{tr}(\Sigma^{-1} \mathbf{K}_p) \end{pmatrix} + O(n_T^{-3/2}).$$

Since $n_T = O_e(k)$, the theorem follows.

References

- Botman, S.L., Moore, T.F., Moriarity, C.L. and Parsons, V.L. (2000). Design and estimation for the National Health Interview Survey, 1995-2004. *Vital and Health Statistics*, 2, 130.
- Cox, D.R., and Snell, E.J. (1968). A general distribution of residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, 30, 248-275.
- Ghosh, M., and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, 91, 95-112.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.
- Gelman, A., and Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 457-511.

- Godambe, V.P., and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). *Journal of Statistical Planning and Inference*, 22, 137-152.
- Morris, C. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10, 65-80.
- Morris, C. (1983). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 11, 515-529.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Sarkar, S., and Ghosh, M. (1998). Empirical Bayes estimation of local area means for NEF-QVF superpopulations. *Sankhyā, Series B*, 60, 464-487.
- Tanner, M.A. (1996). *Tools for Statistical Inference*. New York: Springer.

A nonparametric test for residual seasonality

Tucker McElroy and Scott Holan¹

Abstract

Peaks in the spectrum of a stationary process are indicative of the presence of stochastic periodic phenomena, such as a stochastic seasonal effect. This work proposes to measure and test for the presence of such spectral peaks via assessing their aggregate slope and convexity. Our method is developed nonparametrically, and thus may be useful during a preliminary analysis of a series. The technique is also useful for detecting the presence of residual seasonality in seasonally adjusted data. The diagnostic is investigated through simulation and an extensive case study using data from the U.S. Census Bureau and the Organization for Economic Co-operation and Development (OECD).

Key Words: Multiple testing; Nonparametric density estimation; Seasonal adjustment; Spectral density.

1. Introduction

The presence of a peak in the spectrum of a stationary process is indicative of periodic behavior, such as seasonality or a trading day effect. There is a widespread interest in the identification of such peaks in the engineering and econometrics literature, since a pronounced spectral node will exert a potent influence on the dynamics of the stochastic process. A peak indicates a range of frequencies that offer a relatively large contribution to the overall variance of the stochastic process. If the strength of the peak, assessed through its height and width relative to neighboring values, is sufficiently pronounced, any model of the dynamics that ignores the corresponding periodicities will be misspecified. In both engineering and econometrics, one may be interested in signal extraction or forecasting, both of which are sensitive to the presence of spectral peaks.

By a spectral peak, we refer to a region of the spectral density that has greater spectral mass than its immediate neighbors; a more precise definition is developed below. Due to the applications that we have in mind, our peaks have finite height, and thus correspond to stochastic periodic effects in a stationary process. Thus, we are not principally concerned with the detection of fixed (deterministic) periodic effects, nor with nonstationary periodic phenomena (though we make some extensions to this case in Section 3.4 below), as both of these correspond to a spectral peak with infinite height. The vast literature dealing with the detection of fixed effects is discussed in Priestley (1981); for our applications the periodic aspects of the data are not fixed, but instead evolve over time.

In this paper we focus on the application to seasonal adjustment. Specifically, we concentrate on so-called seasonal peaks, which may occur at the seasonal frequencies (assuming a monthly sampling interval) $\pi/6$, $2\pi/6$, $3\pi/6$,

$4\pi/6$, $5\pi/6$, and $6\pi/6$. The detection of seasonality and residual seasonality presents an important practical problem in federal statistics, and the spectrum is a natural tool towards this end. The frequency domain approach to the detection and analysis of seasonality enjoys wide popularity, because it provides a very natural way to view quasi-periodic behavior. In fact, seasonality is – informally speaking – characterized by the presence of at least one seasonal peak in the spectrum (Nerlove 1964). Frequency domain methods are now employed in X-12-ARIMA (Findley, Monsell, Bell, Otto and Chen 1998) and are part of TRAMO-SEATS (Maravall and Caporello 2004), the two most widely-used seasonal adjustment programs available to the public. Note that frequency domain methods can be implemented via either a parametric (*i.e.*, model-based) or nonparametric approach. We develop a nonparametric diagnostic, which can be invoked to determine the efficacy of *any* seasonal adjustment procedure, either model-based or nonparametric. As noted in Findley, Monsell, Bell, Otto, and Chen (1998), the use of fixed periodic functions alone to model seasonality is typically inadequate for economic data (also see the discussion in Bell and Hillmer 1984).

Spectral peaks at seasonal frequencies in a seasonally adjusted series may indicate inadequacy of the seasonal filters – see Soukup and Findley (1999) for a discussion. At a minimum, seasonal adjustment filters should remove *nonstationary* seasonality and any fixed periodic effects – those phenomena in the observed series that contribute a seasonal pole to the spectrum. However, there is a consensus among seasonal adjusters that it is also desirable to remove some aspects of the *stationary* seasonality as well – hence the explosion of effort in developing model-based seasonal adjustment filters (Bell and Hillmer 1984).

1. Tucker McElroy, Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100. E-mail: tucker.s.mcelroy@census.gov; Scott Holan, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100. E-mail: holans@missouri.edu.

The most important prior literature on this topic is Soukup and Findley (1999), which proposes using an autoregressive spectrum to find “visually significant” peaks – essentially the value of the spectrum at each seasonal frequency (or trading day frequency) is compared to its nearest neighbors, and is classified as a peak if the discrepancy is suitably large. This method is currently implemented in the X-12-ARIMA program from the U.S. Census Bureau (2002). One limitation of this approach is that it has really no statistical component: the significance is not statistical – *i.e.*, it is not associated with a hypothesis test – and the thresholds to determine “visual significance” are determined in an *ad hoc* fashion. This paper provides a statistical significance test for peak detection, and can thus be used to offer supplementary statistical evidence of the presence of a peak.

Another related paper is Newton and Pagano (1983), which develops consistent estimators for the local maximizers of the spectrum. Our approach is slightly different, in that we already know the frequencies of interest (the six seasonal frequencies) but seek to test for the presence of a statistically significant peak. Viewing the true spectral density f as a smooth function (this can be quantified through sufficiently rapid decay of the autocovariance function), a peak is a frequency λ_0 such that

$$\dot{f}(\lambda_0) = 0 \quad \ddot{f}(\lambda_0) < 0, \quad (1)$$

where \dot{f} and \ddot{f} denote first and second derivatives. Clearly, the second derivative must be negative *with some significance* in order for the concept to be meaningful. Upon further reflection, it seems that examining the infinitesimal geometry of f at the single point λ_0 is naïve, since any small spike in the side of a monotonic function may satisfy (1) while being dissociated from more intuitive notions of what constitutes a peak. Therefore, we must have negative convexity in a reasonably large neighborhood of λ_0 . This thinking leads to the diagnostic of this paper: aggregate measures of the slope and convexity of the spectral density, appropriately normalized. Mathematically, these will take the form of kernel-smoothed periodogram estimates, but without the bandwidth being dependent on sample size.

In Section 2 we develop the mathematical ideas of this method, illustrated through two carefully chosen choices of kernels. Section 3 shows how statistical estimators can be formulated, and how statistical peak hypotheses can be tested. The methodology is tested in Section 4; simulations provide a finite sample description of the size and power of our test. We further demonstrate the utility of our methods through an extensive case study involving 130 time series from the U.S. Census Bureau and the Organization for Economic Co-operation and Development (OECD). We use some concepts from the multiple testing literature

(Hochberg 1988) to combine tests based on the individual frequencies together into one diagnostic. Section 5 concludes, and all theorems and proofs are left to the Appendix.

2. Measuring the local geometry of the spectrum

We begin by discussing the geometry of the spectral density (or spectrum) of the time series under consideration. The starting point is to consider measures of slope and convexity of the spectrum that are completely deterministic (*cf.* the approach of Newton and Pagano 1983); later in Section 3 we will consider statistical measures. In Section 2.1 we introduce the concepts of slope and convexity measures. The relevancy of these measures to peak identification is discussed in 2.2, while 2.3 provides two simple kernels as explicit examples.

Suppose that, after suitable transformations and differencing if necessary, X_1, X_2, \dots, X_n is a sample from a zero-mean stationary stochastic process. We will use the notation $X = (X_1, X_2, \dots, X_n)'$. The spectral density $f(\lambda)$ is well-defined so long as the autocovariance function $\gamma_f(h)$ is absolutely summable, and is given by

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_f(h) e^{-ih\lambda} \quad (2)$$

with $i = \sqrt{-1}$ and $\lambda \in [-\pi, \pi]$. It follows that the inverse Fourier transform yields

$$\gamma_f(h) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) e^{ih\lambda} d\lambda, \quad (3)$$

a relation that we will use repeatedly in the sequel. Of course this relationship between γ_g and g holds for any integrable function g , not just a spectral density. Furthermore, denoting the Toeplitz matrix associated with γ_g by $\Sigma(g)$, it follows that

$$\Sigma_{jk}(g) = \gamma_g(j - k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) e^{i(j-k)\lambda} d\lambda.$$

Now from (2), f is d times continuously differentiable if $\sum_{h=-\infty}^{\infty} |h|^d |\gamma_f(h)| < \infty$. We assume that f is twice continuously differentiable for the remainder of the paper (this space of functions will be abbreviated as C^2).

2.1 Measures of slope and convexity

The local geometry of a C^2 function can be described through its first and second derivatives; an aggregate measure of these derivatives is obtained by integrating over a band of frequencies. Alternatively, one may integrate against a function A that has compact support over this band, so long as A provides a suitable proxy for integration over the band. We denote this integral via the general device of a functional θ_A , where

$$\theta_A(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} A(\lambda) f(\lambda) d\lambda. \quad (4)$$

The function A will be referred to as the “kernel” of this functional. Hence the aggregate slope and convexity measures are defined by $\theta_A(\dot{f})$ and $\theta_A(\ddot{f})$, where each dot denotes a single derivative. These functionals give a summary measure of slope and convexity of f over some band $[\mu - \beta/2, \mu + \beta/2] \subset [0, \pi]$, and the corresponding kernels will therefore be denoted $A_{\beta, \mu}$. We consider kernels with the following properties: (i) $A_{\beta, \mu}$ is a C^2 function on $[-\pi, \pi]$; (ii) $A_{\beta, \mu}$ is zero outside the band $[\mu - \beta/2, \mu + \beta/2]$; (iii) $A_{\beta, \mu}$ is symmetric about μ on this band; (iv) $\dot{A}_{\beta, \mu}(\mu \pm \beta/2) = 0$. Condition (iii) ensures that the location of the peak in f is not shifted by employing the kernel $A_{\beta, \mu}$. Note that we do not impose that the total integral of $A_{\beta, \mu}$ be unity, because later we will employ a normalization that will automatically account for the total mass of the kernel. Now by (iv) and integration by parts in (4), we obtain

$$\begin{aligned} \theta_{A_{\beta, \mu}}(\dot{f}) &= -\theta_{\dot{A}_{\beta, \mu}}(f) \\ \theta_{A_{\beta, \mu}}(\ddot{f}) &= \theta_{\ddot{A}_{\beta, \mu}}(f). \end{aligned} \quad (5)$$

These formulas are convenient, because they only require a knowledge of f , not its derivatives (assuming that we can compute $\dot{A}_{\beta, \mu}$ and $\ddot{A}_{\beta, \mu}$). Following the extensive literature on kernels in nonparametric regression and spectral density estimation, we can start with an even kernel A defined on the band $[-\pi, \pi]$ that satisfies (i) and $A(\pm\pi) = 0$. Then $A_{\beta, \mu}$ is defined via

$$A_{\beta, \mu}(\lambda) = \frac{2\pi}{\beta} A\left(\frac{2\pi}{\beta}(\lambda - \mu)\right),$$

and is zero outside the band of frequencies $[\mu - \beta/2, \mu + \beta/2]$. Clearly we must impose $\beta \leq 2\mu$ and $\beta \leq 2(\pi - \mu)$, so that $[\mu - \beta/2, \mu + \beta/2] \subset [0, \pi]$; and the kernel $A_{\beta, \mu}$ satisfies conditions (i)–(iv). Note that we cannot construct these types of measures for μ equal to 0 or π . Using a change of variables, we see that

$$\gamma_{A_{\beta, \mu}}(h) = \exp\{ih\mu\} \gamma_A(h\beta/2\pi), \quad (6)$$

so that the effect of β and μ are in some sense separable. Note that we typically evaluate γ_A at non-integer values, so these relations are obtained by extending (3) to non-integer arguments. The fact that $\gamma_{A_{\beta, \mu}}$ is complex-valued may seem troubling, but actually only its real portion will enter into our statistical estimators. Of course, we are ultimately interested in $\dot{A}_{\beta, \mu}$ and $\ddot{A}_{\beta, \mu}$, which are given by

$$\dot{A}_{\beta, \mu}(\lambda) = \frac{4\pi^2}{\beta^2} \dot{A}\left(\frac{2\pi}{\beta}(\lambda - \mu)\right),$$

and

$$\ddot{A}_{\beta, \mu}(\lambda) = \frac{8\pi^3}{\beta^3} \ddot{A}\left(\frac{2\pi}{\beta}(\lambda - \mu)\right).$$

Later, we will consider the squares of such kernels, and their corresponding inverse Fourier transforms. Hence assuming that $[\mu - \beta/2, \mu + \beta/2] \subset [0, \pi]$, the squares are given by

$$\dot{A}_{\beta, \mu}^2(\lambda) = \frac{16\pi^4}{\beta^4} \dot{A}^2\left(\frac{2\pi}{\beta}(\lambda - \mu)\right)$$

and

$$\ddot{A}_{\beta, \mu}^2(\lambda) = \frac{64\pi^6}{\beta^6} \ddot{A}^2\left(\frac{2\pi}{\beta}(\lambda - \mu)\right).$$

Finally, we notice from (4) that we can rewrite $\theta_A(f)$ as

$$\theta_A(f) = \sum_{h=-\infty}^{\infty} \gamma_A(h) \gamma_f(h). \quad (7)$$

Thus it may be advantageous to determine the $\gamma_A(h)$ sequence from the kernel A . Taking the inverse Fourier Transform of the above slope and convexity kernels, we can construct $\Sigma(\dot{A}_{\beta, \mu})$, $\Sigma(\ddot{A}_{\beta, \mu})$, $\Sigma(\dot{A}_{\beta, \mu}^2)$ and $\Sigma(\ddot{A}_{\beta, \mu}^2)$, as follows:

$$\begin{aligned} \gamma_{A_{\beta, \mu}}(h) &= \frac{2\pi}{\beta} \exp\{ih\mu\} \gamma_A(h\beta/2\pi), \\ \gamma_{\dot{A}_{\beta, \mu}}(h) &= \frac{4\pi^2}{\beta^2} \exp\{ih\mu\} \gamma_{\dot{A}}(h\beta/2\pi), \\ \gamma_{\dot{A}_{\beta, \mu}^2}(h) &= \frac{8\pi^3}{\beta^3} \exp\{ih\mu\} \gamma_{\dot{A}^2}(h\beta/2\pi), \\ \gamma_{\ddot{A}_{\beta, \mu}^2}(h) &= \frac{32\pi^5}{\beta^5} \exp\{ih\mu\} \gamma_{\ddot{A}^2}(h\beta/2\pi). \end{aligned} \quad (8)$$

Thus, if we have the time-domain information $\gamma_f(h)$ for the process $\{X_t\}$, we can compute slope and convexity measures using (7) given the inverse Fourier transform sequence of the appropriate kernels. Since $\gamma_f(h)$ is a symmetric sequence, we only need to consider the real portion of $\gamma_A(h)$ if it happens to be complex.

2.2 Troughs and peaks

The aggregate measures of spectral slope and convexity previously described provide the building blocks for determinants of the local spectral geometry. Our overall interest is in determining whether a given interval of the spectrum is a peak or a trough (or is monotonic). In the second order geometry of calculus, a local maximum has the defining property that the first derivative is zero and the

second derivative is strictly negative. Obviously this requires looking, sequentially, at a slope measure and a convexity measure, defined over the same band of frequencies.

In order to test for the presence of a peak, the sequential approach can be seen as making inferential statements about $\theta_{A_{\beta, \mu}}(\dot{f})$ and $\theta_{A_{\beta, \mu}}(\ddot{f})$. Note, in making these inferential statements we choose μ ahead of time, according to where in the spectrum we wish to detect a peak (or trough); β is chosen according to which frequencies we wish to exclude, a decision based on how local we wish our viewpoint of the spectrum to be. Then we say that μ is a β -aggregate peak (with respect to A) of the spectrum if

$$\theta_A(\dot{f}) = 0 \quad \text{and} \quad \theta_A(\ddot{f}) < 0.$$

The sequential aspect comes from the idea that we generally determine whether $\theta_A(\dot{f}) = 0$ first, and then determine the convexity; this will become more apparent when we consider statistical testing in Section 3.2. In a similar manner we define a β -aggregate trough when $\theta_A(\ddot{f}) > 0$. In terms of hypothesis testing for a peak, we have

$$\begin{aligned} H_0^{(1)}: \theta_A(\dot{f}) = 0 & \quad \text{vs.} \quad H_a^{(1)}: \theta_A(\dot{f}) \neq 0 \\ H_0^{(2)}: \theta_A(\ddot{f}) = 0 & \quad \text{vs.} \quad H_a^{(2)}: \theta_A(\ddot{f}) < 0. \end{aligned}$$

The unusual aspect of this hypothesis test is that we wish to fail to reject $H_0^{(1)}$ first, and then conditional on this test we want to reject $H_0^{(2)}$ in favor of the alternative $H_a^{(2)}$.

2.3 Examples of kernels

There are a host of kernels that satisfy conditions (i) through (iv); we can simply borrow from the literature on nonparametric density estimation. For example, the Parzen and Tukey-Hanning (TH) kernels (discussed in Priestley 1981) are suitable, whereas the Bartlett and Daniell kernels are inappropriate, since (iv) does not hold. In general, one only needs to use (8) to determine the inverse Fourier transforms. In this section, we consider two examples: Quartic and TH. The advantage of these kernels is that they have easily computable first and second derivatives, and their inverse Fourier transforms can be obtained explicitly.

Example 1: Quartic Kernel

We begin by considering a polynomial kernel of degree four, namely a quartic. Imposing all of the constraints (i) through (iv) yields the following form:

$$\begin{aligned} A(\lambda) &= \frac{15}{8\pi^4} (\lambda^4 - 2\pi^2\lambda^2 + \pi^4), \\ \dot{A}(\lambda) &= \frac{15}{8\pi^4} (4\pi^3 - 4\pi^2\lambda), \text{ and} \\ \ddot{A}(\lambda) &= \frac{15}{8\pi^4} (12\pi^2 - 4\pi^2\lambda). \end{aligned}$$

Taking the inverse Fourier transform of the slope and convexity kernels (and their squares) yields

$$\begin{aligned} \gamma_{\dot{A}}(h) &= \frac{15i}{\pi^5} \left(\frac{\pi^2 \sin \pi h}{h^2} + \frac{3\pi \cos \pi h}{h^3} - \frac{3 \sin \pi h}{h^4} \right), \\ \gamma_{\ddot{A}}(h) &= \frac{15}{\pi^5} \left(\frac{\pi^2 \sin \pi h}{h} + \frac{3\pi \cos \pi h}{h^2} - \frac{3 \sin \pi h}{h^3} \right), \\ \gamma_{\dot{A}^2}(h) &= \frac{225}{\pi^9} \left(-\frac{2\pi^4 \sin \pi h}{h^3} - \frac{18\pi^3 \cos \pi h}{h^4} \right. \\ &\quad \left. + \frac{78\pi^2 \sin \pi h}{h^5} + \frac{180\pi \cos \pi h}{h^6} - \frac{18 \sin \pi h}{h^7} \right), \\ \gamma_{\ddot{A}^2}(h) &= \frac{225}{4\pi^{11}} \left(\frac{\pi^4 \sin \pi h}{h} + \frac{6\pi^3 \cos \pi h}{h^2} \right. \\ &\quad \left. - \frac{24\pi^2 \sin \pi h}{h^3} - \frac{54\pi \cos \pi h}{h^4} + \frac{54 \sin \pi h}{h^5} \right), \end{aligned}$$

to which we apply (8) and obtain

$$\begin{aligned} \gamma_{\dot{A}_{\beta, \mu}}(h) &= \frac{30i}{\beta} \exp\{i h \mu\} \left(\frac{\sin k}{k^2} + \frac{3 \cos k}{k^3} - \frac{3 \sin k}{k^4} \right), \\ \gamma_{\ddot{A}_{\beta, \mu}}(h) &= \frac{30}{\beta^2 \pi} \exp\{i h \mu\} \left(\frac{\sin k}{k} + \frac{3 \cos k}{k^2} - \frac{3 \sin k}{k^3} \right), \\ \gamma_{\dot{A}_{\beta, \mu}^2}(h) &= \frac{1,800\pi}{\beta^3} \exp\{i h \mu\} \left(\frac{2 \sin k}{k^3} + \frac{18 \cos k}{k^4} \right. \\ &\quad \left. - \frac{78 \sin k}{k^5} - \frac{180 \cos k}{k^6} + \frac{180 \sin k}{k^7} \right), \text{ and} \\ \gamma_{\ddot{A}_{\beta, \mu}^2}(h) &= \frac{1,800}{\beta^5 \pi} \exp\{i h \mu\} \left(\frac{\sin k}{k} + \frac{6 \cos k}{k^2} \right. \\ &\quad \left. - \frac{24 \sin k}{k^3} - \frac{54 \cos k}{k^4} + \frac{54 \sin k}{k^5} \right), \end{aligned}$$

where $k = h\beta/2$. Note that $\gamma_{\dot{A}_{\beta, \mu}^2}(0) = 240\pi/(7\beta^3)$ and $\gamma_{\ddot{A}_{\beta, \mu}^2}(0) = 360/(\beta^5\pi)$ follow by application of L'Hopital's rule. These formulas allow us to construct the appropriate Toeplitz matrices for the diagnostic (as discussed in Section 3.1 below, it suffices to consider the real part of these sequences).

Example 2: TH Kernel

A similar shape to the quartic can be obtained through the use of a cosine function. The following choice satisfies all the stated conditions on a kernel:

$$\begin{aligned} A(\lambda) &= \frac{1}{2\pi} (1 + \cos \lambda), \\ \dot{A}(\lambda) &= \frac{1}{2\pi} (-\sin \lambda), \text{ and} \\ \ddot{A}(\lambda) &= \frac{1}{2\pi} (-\cos \lambda). \end{aligned}$$

This function is identical to the Tukey-Hanning lag window, though here we apply it as a spectral window (see Priestley 1981). Hereafter it will be referred to as the TH kernel. Taking the inverse Fourier transform of the slope and convexity kernels (and their squares) yields

$$\begin{aligned}\gamma_{\dot{A}}(h) &= \frac{i}{4\pi^2} \left(\frac{\sin \pi(h+1)}{h+1} - \frac{\sin \pi(h-1)}{h-1} \right), \\ \gamma_{\ddot{A}}(h) &= -\frac{1}{4\pi^2} \left(\frac{\sin \pi(h+1)}{h+1} + \frac{\sin \pi(h-1)}{h-1} \right), \\ \gamma_{\dot{A}^2}(h) &= \frac{1}{16\pi^3} \left(\frac{2\sin \pi h}{h} - \frac{\sin \pi(h+2)}{h+2} - \frac{\sin \pi(h-2)}{h-2} \right), \text{ and} \\ \gamma_{\ddot{A}^2}(h) &= \frac{1}{16\pi^3} \left(\frac{2\sin \pi h}{h} + \frac{\sin \pi(h+2)}{h+2} + \frac{\sin \pi(h-2)}{h-2} \right).\end{aligned}$$

Now applying (8) yields

$$\begin{aligned}\gamma_{\dot{A}_{\beta,\mu}}(h) &= \frac{i}{2\beta} \exp\{ih\mu\} \left(\frac{\sin(k+\pi)}{k+\pi} - \frac{\sin(k-\pi)}{k-\pi} \right), \\ \gamma_{\ddot{A}_{\beta,\mu}}(h) &= -\frac{\pi}{\beta^2} \exp\{ih\mu\} \left(\frac{\sin(k+\pi)}{k+\pi} + \frac{\sin(k-\pi)}{k-\pi} \right), \\ \gamma_{\dot{A}_{\beta,\mu}^2}(h) &= \frac{\pi}{2\beta^3} \exp\{ih\mu\} \left(\frac{2\sin k}{k} - \frac{\sin(k+2\pi)}{k+2\pi} - \frac{\sin(k-2\pi)}{k-2\pi} \right), \text{ and} \\ \gamma_{\ddot{A}_{\beta,\mu}^2}(h) &= \frac{2\pi^3}{\beta^5} \exp\{ih\mu\} \left(\frac{2\sin k}{k} - \frac{\sin(k+2\pi)}{k+2\pi} + \frac{\sin(k-2\pi)}{k-2\pi} \right),\end{aligned}$$

where $k = h\beta/2$. Note that $\gamma_{\dot{A}_{\beta,\mu}^2}(0) = \pi/\beta^3$ and $\gamma_{\ddot{A}_{\beta,\mu}^2}(0) = 4\pi^3/\beta^5$ follow by application of L'Hopital's rule⁴ (using the convention that $\sin(0)/0 = 1$). These formulas allow us to construct the appropriate Toeplitz matrices for the diagnostic (again, as discussed in Section 3.1 below, it suffices to consider the real part of these sequences).

3. Statistical methodology

Of course we do not typically have knowledge of the spectrum f , and thus it is usually necessary to form estimates from the data. In this section we describe statistical estimates of slope and convexity measures that are consistent and simple to compute in the time-domain. Under some mild additional assumptions, these estimates are asymptotically normal, which will be advantageous when performing hypothesis tests. In Section 3.1 the statistical estimates are defined, and their asymptotic properties are

discussed. Section 3.2 discusses the application to peak testing, and 3.3 gives an extension to joint peak testing, which facilitates an important application in seasonal adjustment. Section 3.4 discusses extensions to trend nonstationary data.

3.1 Estimators of slope and convexity

We begin by noting that the quadratic form (for any integrate function g)

$$\frac{1}{n} X' \Sigma(g) X = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) I(\lambda) d\lambda,$$

where I denotes the periodogram. Although the periodogram is typically defined at the Fourier frequencies ($2\pi j/n$; $j = 1, \dots, \lfloor n/2 \rfloor$) we define it at a continuous band of frequencies as follows

$$\begin{aligned}I(\lambda) &= \frac{1}{n} \left| \sum_{t=1}^n X_t e^{-i\lambda t} \right|^2 \\ &= \sum_{h=-n}^{n-1} R(h) e^{-i\lambda h}, \lambda \in [-\pi, \pi] \quad (9)\end{aligned}$$

with $R(h)$ equal to the sample (uncentered) autocovariance function. This gives an elegant way of passing from the time-domain to the frequency-domain, and is well-known in the time series literature (see Taniguchi and Kakizawa 2000). Moreover, such integrals of the periodogram are generally consistent, *i.e.*, $\theta_g(I) \xrightarrow{a.s.} \theta_g(f)$ as $n \rightarrow \infty$, under mild conditions discussed below (note that the inconsistency of the periodogram is resolved by the spectral aggregation against the function g , as shown in the Appendix). Therefore, we obtain statistical estimates of the slope and convexity measures f by using a “plug in” approach, *i.e.*, we simply replace f by I in $\theta_{A_{\beta,\mu}}$. In particular,

$$\begin{aligned}\hat{\theta}_{\dot{A}_{\beta,\mu}}(\dot{f}) &= -\theta_{\dot{A}_{\beta,\mu}}(I) = -\frac{1}{n} X' \Sigma(\dot{A}_{\beta,\mu}) X, \text{ and} \\ \hat{\theta}_{\ddot{A}_{\beta,\mu}}(\ddot{f}) &= \theta_{\ddot{A}_{\beta,\mu}}(I) = \frac{1}{n} X' \Sigma(\ddot{A}_{\beta,\mu}) X. \quad (10)\end{aligned}$$

This definition makes use of (5), which accounts for the minus sign in the slope measure. In order to compute the estimate, we utilize the time-domain representation (expressed as a quadratic form). This representation is convenient in that we only need determine a suitable length of the sequences $\gamma_{\dot{A}_{\beta,\mu}}(h)$ and $\gamma_{\ddot{A}_{\beta,\mu}}(h)$, form the Toeplitz matrices $\Sigma(\dot{A}_{\beta,\mu})$ and $\Sigma(\ddot{A}_{\beta,\mu})$, and then compute the quadratic forms. Note that the inverse Fourier transforms of $\dot{A}_{\beta,\mu}$ and $\ddot{A}_{\beta,\mu}$ need only be determined once (see Section 2.3 for some explicit examples) and can be done ahead of time, and then applied repeatedly to many different time series.

In order to compute the time domain representation of the slope and convexity measures in (10), we utilize (8),

e.g., see the formulas in Examples 1 and 2. Of course, this will in general result in $\Sigma(\dot{A}_{\beta,\mu})$ and $\Sigma(\ddot{A}_{\beta,\mu})$ being complex. However, even if $\Sigma(g)$ (where g can be $A_{\beta,\mu}$, $\dot{A}_{\beta,\mu}$, or $\ddot{A}_{\beta,\mu}$) is a complex Toeplitz matrix, $X'\Sigma(g)X$ will always be real. From (8), it is easy to see that $\Sigma(g) = M + iN$ where M is real, symmetric, and Toeplitz, and N is real, skew-symmetric, and Toeplitz. Hence $X'NX = 0$ for any vector X , so that $X'\Sigma(g)X = X'MX$. Therefore, for the purposes of computing the statistical slope and convexity measures, we may take the real part of $\gamma_A(h)$ in (8).

Not only are these statistical estimates consistent, they are also asymptotically normal under some additional conditions (discussed in the Appendix). However, in order to construct a suitable normalization it will be necessary to estimate their variation. The asymptotic variance of $\theta_g(I)$ is $\theta_{g^2}(f^2)$ (if g is supported on $[0, \pi]$), which can be consistently estimated via $\theta_{g^2}(I^2)/2$. (The factor of 2 is required, since the integral of I^2 tends to the corresponding integral of $2f^2$ —see Chiu (1988)). This can be given a time-domain representation as follows. Let $R = \{R(1-n), \dots, R(0), \dots, R(n-1)\}'$ be a $2n-1$ vector of sample autocovariances, and let $\Sigma(g^2)$ be $2n-1$ dimensional in the following formula: $R'\Sigma(g^2)R/2 = \theta_{g^2}(I^2)/2$. This relationship can be easily verified using (9). Thus we will normalize $\theta_g(I)$ by the square root of $\theta_{g^2}(I^2)/2$. Hence our normalized statistical measures of slope and convexity are given by

$$-\psi_{\dot{A}_{\beta,\mu}}(I) = -\frac{\theta_{\dot{A}_{\beta,\mu}}(I)}{\sqrt{\theta_{\dot{A}_{\beta,\mu}^2}(I^2)/2}} = -\frac{1}{n} \frac{X'\Sigma(\dot{A}_{\beta,\mu})X}{\sqrt{R'\Sigma(\dot{A}_{\beta,\mu}^2)R/2}}$$

and

$$\psi_{\ddot{A}_{\beta,\mu}}(I) = -\frac{\theta_{\ddot{A}_{\beta,\mu}}(I)}{\sqrt{\theta_{\ddot{A}_{\beta,\mu}^2}(I^2)/2}} = \frac{1}{n} \frac{X'\Sigma(\ddot{A}_{\beta,\mu})X}{\sqrt{R'\Sigma(\ddot{A}_{\beta,\mu}^2)R/2}},$$

where the dimensions of the Σ matrices are either n or $2n-1$ as appropriate. The asymptotic properties of $\psi_{\dot{A}_{\beta,\mu}}(I)$ and $\psi_{\ddot{A}_{\beta,\mu}}(I)$ are discussed in the Appendix. In summary, both $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$ and $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$ are marginally asymptotically $N(0, 1)$ under $H_0^{(1)}$ and $H_0^{(2)}$ respectively and the assumptions discussed in the Appendix. Simulations indicate that the variance normalization is slow to converge, and its correlation with numerator causes a degree of non-normality in smaller samples. Based on the histogram of the distribution simulated under a Gaussian white noise Null hypothesis with $n = 360$ and 10,000 replications (Figure 1) there is close agreement to the normal distribution, except at the extremes in the tails. Section 4 explores this behavior further through simulation studies.

3.2 Applications to single peak testing

We now consider the application to peak testing. Recall that we have an initial Null Hypothesis $H_0^{(1)}$ that we must fail to reject in order to proceed. This can be interpreted as saying there is insufficient evidence to conclude that the first derivative (slope) of the spectral density is significantly different from zero. Now we know that $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$ is asymptotically $N(0, 1)$ under $H_0^{(1)}$ and the assumptions discussed in the Appendix. If we further suppose that a sufficiently small value x is obtained for the test statistic, we will not be able to reject $H_0^{(1)}$ with any confidence. In that case, we can consider the hypothesis $H_0^{(2)}$, which we seek to reject; this is tested via $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$. Although $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$ and $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$ are asymptotically correlated (see Theorem 1 of the Appendix) we will consider the slope and convexity tests as if they were done separately (this correlation can be estimated, and used to determine the distribution of the convexity diagnostic conditional on the slope diagnostic; however, the interpretation of p -values becomes muddled. For simplicity, we treat the tests separately, one at a time, and do not explicitly account for the correlation). Our testing procedure is then conducted as follows:

1. Perform the 2-sided test of $H_0^{(1)}$ using $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$.
2. Let p be the p -value associated with the first test statistic's value $x = -\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$, with x and p related by $p = 2\Phi(-|x|)$.
3. If $p > 0.05$ (or some other pre-determined tolerance level) proceed; else conclude that there is no peak present.
4. Perform the lower 1-sided test of $H_0^{(2)}$ using $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$.
5. Reject $H_0^{(2)}$ and conclude that there is a peak if $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I) < \Phi^{-1}(\alpha)$, where α is the level of the convexity test.

3.3 Joint peak testing: Application to seasonal adjustment

We now consider the situation where we wish to test for several spectral peaks simultaneously. Clearly we could design a kernel with several nodes, one at each peak, but this would merely be the sum of several individual spectral peak diagnostics. It would have the disadvantage that a significant spectral peak in one place could cancel a significant spectral trough elsewhere. Therefore, we would prefer a test that examines a set of spectral diagnostics within a multiple testing paradigm.

For example, consider the context of testing for spectral peaks in seasonally adjusted data. There are six seasonal peaks of interest, but we must restrict attention to five due to aliasing problems (the peak at frequency π cannot be identified). If one or more of the spectral peaks is significant, we must reject our seasonal adjustment procedure (since it has failed to remove all of the peaks); therefore, we are in a multiple testing situation, and will utilize a method that controls the familywise error rate (FWER) proposed by Hochberg (1988) and described in Benjamini and Hochberg (page 294, 1995). Restricting attention to the issue of convexity, we have Null Hypotheses $H_{(i)}^{(2)}$ for each of the five seasonal frequencies. In our setting, the procedure of Hochberg (1988) is to compute p -values for the convexity test at each of the five seasonal frequencies, and order them as $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq p_{(4)} \leq p_{(5)}$, with corresponding Null Hypotheses denoted by $H_{(i)}$. For a specified FWER of level α (e.g., $\alpha = 0.05$), let k be the largest i for which $p_{(i)} \leq i/(6-i)\alpha$; then reject all $H_{(i)}$ for $i \leq k$.

When using such a procedure, we should make Type I errors – i.e., identifying at least one seasonal frequency as having negative convexity when none is present – roughly α proportion of the time (if we were to restrict attention to $H_0^{(2)}$, the convexity hypothesis). The advantage of the Hochberg familywise error rate approach (H-FWER) is that it dramatically improves the statistical power compared to other methods. The validity of this method requires independence of the test statistics under consideration, and so for this reason we take five kernels A_1, \dots, A_5 – centered at the seasonal frequencies $\pi/6, \dots, 5\pi/6$ respectively – that have disjoint support. Then Theorem 1 can be generalized to obtain asymptotic independence of the five convexity test statistics (see the discussion after Theorem 1 in the Appendix). Of course, we also conduct five separate tests of the slope at each seasonal frequency, where we must fail to reject in each case in order to proceed.

As a final remark, we note that in practice a seasonal adjustment is rarely rejected on the basis of significant spectral mass at the fifth seasonal frequency of $5\pi/6$ (Findley 2006). This is partly due to the difficulty in assigning an interpretation to this frequency. Therefore, practitioners may be more interested in a “four-peak test” that focuses on the first four seasonal frequencies; one obtains this test by an obvious modification of the H-FWER procedure described above.

3.4 Extending to nonstationary data

The methodology given above assumes that the data are a sample from a stationary process. However, in the context of seasonal adjustment, it is usually the case that the seasonally adjusted data are once or twice integrated. In this case one would difference the seasonally adjusted data once

or twice, and then apply the diagnostics. Now application of the differencing operators $1 - B$ and $(1 - B)^2$ are essentially high-pass filters, which can be expected to attenuate residual spectral peaks close to frequency zero (in particular, the first seasonal frequency at $\pi/6$). Thus it may be desirable to apply the diagnostic to the pseudo-spectrum instead; this can be done if the support of the kernel is bounded away from the poles in the spectral density.

Suppose that the observed data are now Y_{1-d}, \dots, Y_n for d the order of trend differencing (so usually $d = 1$ or 2). When the observed data are differenced, we obtain the sample X , which is strictly stationary. The pseudo-spectral density of the $\{Y_t\}$ process is $g(\lambda) = f(\lambda) |1 - e^{-i\lambda}|^{-2d}$, where f is the spectrum of $\{X_t\}$. This pseudo-spectrum could be estimated via $\hat{g}(\lambda) = I(\lambda) |1 - e^{-i\lambda}|^{-2d}$, where I is the periodogram of X as before; this is the re-coloring approach of Nerlove (1964). Then $\theta_A(g)$ is well-defined so long as $A(\lambda) |1 - e^{-i\lambda}|^{-2d}$ is an integrable function; essentially we must ensure that frequency zero is excluded from the support of the kernel A . Since A is centered around seasonal frequencies in practice, we can easily contrive this condition. The corresponding estimator is then

$$\hat{\theta}_A(\hat{g}) = \theta_{\tilde{A}b}(I),$$

where $b(\lambda) = |1 - e^{-i\lambda}|^{-2d}$. The estimator is well-defined if $\tilde{A}b$ is integrable; moreover the asymptotic properties discussed in the Appendix for the stationary case extend to this case as well, so long as $\tilde{A}b$ is bounded.

This extension may be more appealing to some researchers. However, the cost is that the inverse Fourier transform of $\tilde{A}b$ must be determined, which requires some additional mathematical work. In the simulation studies and data illustrations in Section 4 we trend difference the seasonally adjusted data, but do not implement the correction factor b in the kernel.

4. Empirical studies

Having developed the theoretical aspects of the spectral diagnostic, we now turn to its performance in practice. We first present some results obtained from simulation, which provide insight into the size and power properties of the test statistic in finite samples. Then we investigate the size and power empirically, by applying the spectral diagnostics to a suite of 130 time series (65 U.S. Census Bureau series and 65 OECD series); we consider both the original and the seasonally adjusted series, and make comparisons to the Visual Significance, M7 and M8 quality control diagnostics of X-12-ARIMA (U.S. Census Bureau 2002). Additional empirical studies can be found in Evans, Holan and McElroy (2006).

4.1 Simulation study

To evaluate the performance of our diagnostics we conducted several simulations. The first set of simulations examines size (level) for the single peak diagnostic. For this simulation we considered the slope and convexity diagnostics separately. Although in practice, when considering the slope diagnostic, we wish to fail to reject the Null hypothesis $H_0^{(1)}$, here we are interested in empirically investigating the distributional properties, and so we impose the usual definition of size for this study. So we simulated Gaussian white noise which satisfies the assumptions of Theorem 1 as well as satisfying $\psi_A(I) = \psi_{\hat{A}}(I) = 0$, so that $H_0^{(1)}$ and $H_0^{(2)}$ are true. Of course, there are many processes for which both $H_0^{(1)}$ and $H_0^{(2)}$ are true simultaneously – for example, any process with locally flat spectral density; however, due to asymptotic considerations it suffices to consider white noise. For a (large) sample size of $n = 360$, using the TH kernel with $\mu = \pi/6$ and $\beta = \pi/6$ (this corresponds to a kernel centered on the interval $[0, \pi/6]$), 10,000 repetitions yields an empirical distribution of the normalized diagnostics, $\psi_A(I)$ and $\psi_{\hat{A}}(I)$, whose histograms are displayed in Figure 1. Henceforth, let δ and α denote the levels associated with the slope and convexity tests respectively. Note that in this case we define level to mean the probability of rejecting $H_0^{(j)}$ ($j = 1, 2$) when $H_0^{(j)}$ is true. Although, in practice, in the case of the slope hypothesis we wish to fail to reject we follow the strict definition of level and assume (for the purposes of this simulation) that the null hypothesis $H_0^{(1)}$ for the slope holds true. Similarly the null hypothesis for the convexity is $H_0^{(2)}$. Both the slope and convexity hypotheses are evaluated independently. Table 1 summarizes the results using both kernels from Section 2, for various sample sizes; the indicated δ, α -levels are for the nominal 5 percent level. Additionally, other choices of μ and β (not shown here) yielded similar results. As depicted in this study, in smaller samples, we observed skewness in the distribution which seems to be due to correlation between $\theta_A(I)$ and $\theta_{\hat{A}}(I^2)$. Also, note that the size for the convexity test is larger for the quartic kernel than for the TH kernel.

Next we consider the empirical power for our single peak diagnostic. In this setting we evaluate the power based on a joint test of the slope and convexity. Specifically, we wish to fail to reject $H_0^{(1)}$ while simultaneously rejecting $H_0^{(2)}$, at $\delta = \alpha = 0.05$, and thus correctly identify spectral peaks. Since our composite Null hypothesis is that there is no peak, the Alternative hypothesis includes processes such as the $AR(2)$ given by

$$(1 - 2\rho\cos\omega B + \rho^2 B^2)X_t = \varepsilon_t \quad (11)$$

with white noise variance σ^2 , associated with some fixed frequency $\omega \in [0, \pi]$. The spectrum associated with the

process in (11) is given by $f(\lambda) = \sigma^2[1 - 2\rho\cos\omega e^{-i\lambda} + \rho^2 e^{-2i\lambda}]^{-2}$, which is maximized at $\lambda_0 = \cos^{-1}(\cos\omega(1 + \rho^2)/2\rho)$. Therefore one can explore the power of a peak-testing procedure by simulating from (11) with various choices of ρ , ω , and σ . Table 2 presents the result of 10,000 simulations, of various sample sizes, from the $AR(2)$ cycle model given in (11) with peak at $\mu = \pi/6$ and bandwidth set at $\beta = \pi/6$. The peak strength is parametrized through ρ , which we vary from 0.85 to 0.95; clearly $H_0^{(1)}$ and $H_a^{(2)}$ are both true for this model. In other words, there are spectral peaks, of different heights, at $\lambda = \pi/6$. This AR cycle model was chosen because it provides a convenient parametrization of spectral peak location and shape. Additionally, this choice of β is compatible with the seasonal adjustment setting, as this provides the maximum window width while avoiding overlapping spectral peaks. As expected, the power of our diagnostic increases with sample size and peakedness, ranging from 0.227 (quartic kernel) in small samples having weak spectral peak to ≈ 0.95 (TH kernel) in larger samples having a more pronounced spectral peak (see Table 2). Note that in this procedure the innovation variance is set equal to one, but it is immaterial due to the normalization of the diagnostic. In summary, both the quartic and TH kernels possess decent size and power properties. Generally, the quartic kernel seems to have superior size and power, so it would be preferable for spectra of this form (note that the lower power of the TH kernel is in part due to its being undersized). Additionally, it seems that smaller values of β (results not shown) require a greater sample size; a smaller β corresponds to a more refined “viewing” of the spectral peak, which would require more data to handle the resolution.

Although the individual peak testing scenario provides the foundation for our joint testing framework, as noted, the joint testing framework provides important methodology for applications in federal statistics. The application of importance is the evaluation of effective seasonal adjustment through the exploration of residual seasonality. Thus, it is of particular interest to know how our multiple testing approach performs in simulation. Therefore, in order to investigate the size and power associated with our joint test, we simulated 10,000 repetitions from a Gaussian white noise process and from an $AR(25)$ model obtained as a fit to the Current Employment Series (Employed Males, aged 16 to 19). Our goal in the power study was to construct an $AR(p)$ process (because of its ease in simulation and desirable theoretical properties as a parametric spectral estimator – see Parzen 1983) with (stationary) spectral peaks that are realistic, or close to what might be found in practice. Thus we obtain our $AR(25)$ model – fitted via maximum likelihood using AIC – which has similar seasonal dynamics (local spectral behavior) to the Current Employment Series (CES).

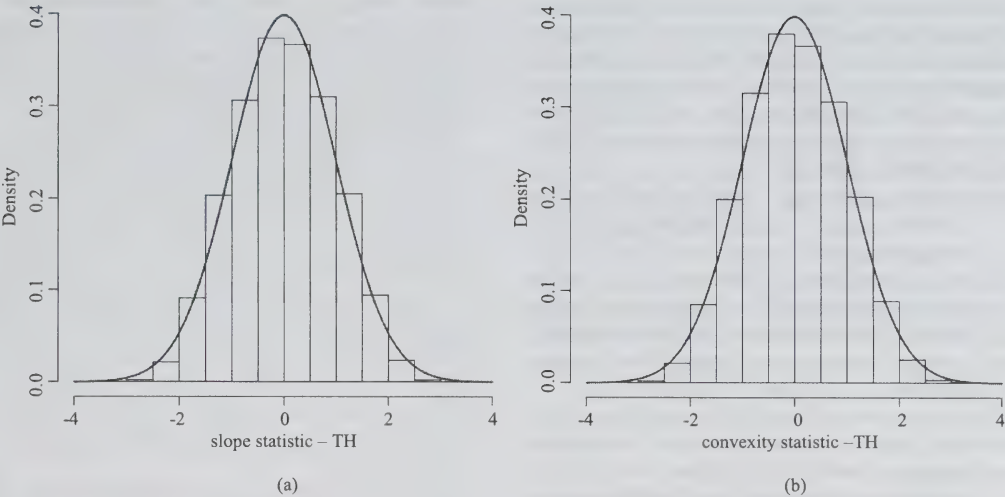


Figure 1 Histogram of distribution of $-\sqrt{n}\psi_{A_{p,n}}(J)$ (a) and $\sqrt{n}\psi_{A_{p,n}}(J)$ (b) under a Gaussian white noise Null hypothesis using the TH kernel. The sample size is $n = 360$ with 10,000 replications

Table 1 Results of size simulation for the single peak diagnostic. Here $\mu = \beta = \pi/6$ and 10,000 repetitions were used. The slope and convexity diagnostics were investigated separately for both the quartic and TH kernels

Size for Single Peak $\mu = \beta = \pi/6$												
Slope							Convexity					
Quartic Kernel				TH Kernel			Quartic Kernel			TH Kernel		
n	Mean	Stdev	δ -level	Mean	Stdev	δ -level	Mean	Stdev	α -level	Mean	Stdev	α -level
120	0.003	0.903	0.007	-0.011	0.903	0.008	-0.065	0.852	0.032	0.025	0.888	0.018
144	-0.004	0.920	0.014	-0.011	0.927	0.015	-0.077	0.882	0.042	0.006	0.892	0.025
180	-0.003	0.942	0.022	0.002	0.920	0.017	-0.071	0.892	0.043	0.005	0.902	0.028
288	0.003	0.954	0.027	-0.002	0.950	0.025	-0.072	0.921	0.051	-0.006	0.926	0.033
360	0.003	0.962	0.032	-0.009	0.954	0.031	-0.056	0.922	0.051	0.006	0.951	0.040

Table 2 Results of power simulation for the single peak diagnostic. Here $\mu = \beta = \pi/6$ and 10,000 repetitions were used. The alternative hypothesis is given by the $AR(2)$ model defined by (11). The slope and convexity diagnostics were investigated simultaneously for both the quartic and TH kernels using $\delta = \alpha = 0.05$ for both tests (see Section 4.1)

Power for Single Peak $\mu = \beta = \pi/6 - (\delta, \alpha) = (0.05, 0.05)$						
Quartic Kernel				TH Kernel		
n	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$
120	0.227	0.438	0.758	0.147	0.335	0.670
144	0.287	0.532	0.856	0.208	0.431	0.799
180	0.354	0.643	0.923	0.272	0.567	0.901
288	0.447	0.755	0.949	0.372	0.706	0.950
360	0.601	0.872	0.937	0.537	0.859	0.948

To evaluate size we considered both a test based on convexity alone and a test based on the slope and convexity simultaneously. The tests based on convexity alone (C) were performed at the nominal α -levels of 0.05 and 0.10, using the H-FWER method to control the FWER. The tests based on both the slope and convexity simultaneously (S, C) were performed as follows:

1. Perform multiple tests of convexity, $H_0^{(2)}$, using the H-FWER method to control the FWER at level α (which is either 0.05 or 0.10).
2. For any peaks found significant in Step 1 perform individual slope test, $H_0^{(1)}$, at level δ (which is either 0.10 or 0.25). Note here we wish to fail to reject $H_0^{(1)}$ in order to declare any “peaks” as statistically significant.
3. Declare there is a statistically significant peak if Step 1 finds any seasonal frequency with significant aggregate convexity in the spectrum, and if Step 2 simultaneously fails to find any significant aggregate slope for the corresponding seasonal frequency.

The results of this simulation are summarized in Table 3. One aspect of this procedure that deserves further explanation is Step 2 where δ (the level for the slope test) is taken equal to 0.10 and 0.25. Although the slope testing aspect of the procedure is conducted on an individual peak basis, it seems reasonable to try and be conservative. The issue here is that even if some of the individual slope tests are rejected we may still proceed in other cases. Thus the

situation encountered here differs from the classical “no peaks” hypothesis which can be rejected if a single peak is found. Of course, since we are conducting each slope hypothesis test on an individual peak basis, any δ -level greater than 0.05 would be considered more conservative.

While we cannot expect the combined procedure (S, C) to have size approaching the nominal (because using the slope test throws off the Type I error rate), neither is the size highly accurate in the case of using the convexity test alone (C), as can be seen by examining the $\alpha = 0.10$ case with $n = 288, 360$. Here the convexity for the quartic kernel is over-sized, whereas in the single peak case the convexity test has accurate size (Table 1) for these sample sizes. Note that H-FWER only produces an approximately correctly sized procedure; another factor is that the five peak tests are only asymptotically independent. It is for these reasons that the empirical size found in Table 3 differ somewhat from the nominal levels.

To investigate power we considered the same 3 step procedure as outlined above. However, for this simulation we only considered joint slope - convexity testing and examined four pairs of (δ, α) levels; $(\delta, \alpha) = (0.10, 0.05), (0.25, 0.05), (0.10, 0.10)$ and $(0.25, 0.10)$. The results of this simulation (Table 4) indicate tremendous power even at sample sizes as small as $n = 120$. This is extremely important as $n = 120$ is representative of the size samples encountered in practice when conducting seasonal adjustment (*i.e.*, 10 years of monthly data). For samples sizes $n = 144$, greater than 90% power was achieved.

Table 3

Results of size simulation for the multiple peak diagnostic. Here 10,000 repetitions were used. The convexity test was investigated separately with the FWER controlled at $\alpha = 0.05$ and $\alpha = 0.10$ using the H-FWER method. Additionally, the slope and convexity diagnostics were investigated simultaneously using the H-FWER method for the convexity controlling the FWER at $\alpha = 0.05$ and $\alpha = 0.10$, while the slope was evaluated at $\delta = 0.25$ and $\delta = 0.10$. Both the quartic and TH kernels were used for both tests (see Section 4.1)

<i>n</i>	C = 0.05		(S, C) = (0.10, 0.05)		Size for Multiple Peaks H-FWER (S, C) = (0.25, 0.05)		C = 0.10		(S, C) = (0.10, 0.10)		(S, C) = (0.25, 0.10)	
	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH
120	0.006	0.002	0.006	0.002	0.005	0.002	0.076	0.047	0.070	0.044	0.076	0.046
144	0.009	0.002	0.011	0.002	0.008	0.003	0.087	0.053	0.090	0.051	0.086	0.050
180	0.019	0.005	0.020	0.006	0.017	0.005	0.107	0.062	0.097	0.059	0.093	0.057
288	0.031	0.009	0.026	0.008	0.025	0.008	0.117	0.069	0.116	0.069	0.112	0.068
360	0.042	0.012	0.045	0.019	0.035	0.015	0.140	0.087	0.133	0.084	0.129	0.087

Table 4

Results of power simulation for the multiple peak diagnostic. Here 10,000 repetitions were used. The slope and convexity diagnostics were investigated simultaneously using the H-FWER method for the convexity controlling the FWER at $\alpha = 0.05$ and $\alpha = 0.10$. For the slope $\delta = 0.25$ and $\delta = 0.10$. Both the quartic and TH kernels were used for both tests (see Section 4.1)

<i>n</i>	(0.10, 0.05)		Power for Multiple Peaks H-FWER (0.25, 0.05)		(0.10, 0.10)		(0.25, 0.10)	
	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH
120	0.877	0.897	0.860	0.881	0.997	0.997	0.996	0.998
144	0.943	0.952	0.942	0.950	0.999	1.00	0.999	1.00
180	0.989	0.992	0.989	0.992	1.00	1.00	0.999	1.00
288	1.00	1.00	0.999	0.999	1.00	1.00	0.999	1.00
360	1.00	1.00	0.998	0.999	1.00	1.00	0.998	0.999

4.2 Case studies

We also considered 130 time series, 65 from the U.S. Census Bureau and 65 from OECD. These series consist of 35 U.S. Manufacturing series, 10 U.S. Housing series, 10 U.S. Import/Export series, and 10 U.S. Retail series; there are also 22 German series, 15 Euro-area series, 11 French series, and 17 Great Britain series from OECD, covering the sectors of manufacturing, retail, wholesale, foreign trade, unemployment, and industry. For every series, we computed the seasonal peak tests for both the raw data (logged) and the seasonally adjusted data (logged) – using the $x11$ specification of $X-12-ARIMA$ – with both the quartic and TH kernels. We employed the H-FWER procedure controlling the FWER at $\alpha = 0.05, 0.10$, and where the threshold for the slope tests was $\delta = 0.25$ and $\delta = 0.10$ at each peak (see Section 4.1). Note that $\delta = 0.25$ and $\delta = 0.10$ produced similar results and thus, for the sake of brevity, only results for $\delta = 0.10$ are presented here. The results for $\delta = 0.25$ are available upon request from the first author. For both the raw and seasonally adjusted data, a single trend difference was used (as is the case for the Visual Significance diagnostic, described below) before applying the seasonal peaks test.

In addition, we present the M7 and M8 statistics as well as the results of the Visual Significance (VS) diagnostic both before and after adjustment. The M7 quality control statistic measures the amount of stable seasonality relative to the moving seasonality in the original series, with values greater than 1 indicating that the seasonality in the series is not identifiable (Lothian and Morry 1978); similarly, the M8 statistic measures the size of the fluctuations in the seasonal component, with a similar interpretation. We also considered the robust nonparametric Kruskal-Wallis test (U.S. Census Bureau 2002) for the presence of seasonality assuming stability. VS is based on an $AR(30)$ spectrum estimate of the raw and seasonally adjusted series, and is described in Soukup and Findley (1999). For Tables 5-10, each cell entry lists which seasonal frequencies were found to have a significant peak, with j corresponding to $\pi j/6$ for $j = 1, 2, 3, 4, 5$; an entry of \emptyset indicates that no peaks were detected. For the M7 and M8 diagnostics, only the value is reported since there is no associated p -value (and they are only pertinent to the raw series).

The results of this empirical study can be found in Tables 5-10. All of the Kruskal-Wallis statistics were significant with $p = 0.000$, so these are not reported in the tables. The set of columns corresponding to the “Original Data” heading can be seen as giving empirical power (for each subset of series), assuming that each series is indeed seasonal and has seasonal spectral peaks. That is, the Total “correct” number gives the proportion of times each method correctly identified seasonality, and hence this proportion is

a crude proxy for empirical power. We also report the average number of peaks that were identified, which is an empirical measure of the efficacy of the methods (the more peaks correctly identified, the better). The set of columns for “SA Data” gives an empirical size (for each subset of series), assuming that seasonal adjustment has indeed removed the spectral peaks. These are rough considerations, since we do not really know a priori whether the SA Data has been adequately adjusted.

The VS identifies all of the raw series as seasonal and most of the SA series as having no spectral seasonal peaks; the M7 and M8 diagnostics perform similarly, though of course they do not indicate which seasonal peaks are present in the raw data. Our procedure indicates a few cases (when $\alpha = 0.10$ for the convexity tests) where the adjustment may be inadequate, but these are within the scope of the expected proportion of Type I errors. For the raw series, the empirical power (*i.e.*, total proportion correct) for our method ranges between 0.66 and 0.89, with higher power for the $\alpha = 0.10$ level, as expected. In many cases the indicated peaks are the same as VS, but sometimes are quite different. Note that the average number of peaks detected for raw series was typically much higher for our procedure over the VS method, which often had an average around 3.2. When the α level was increased from 0.05 to 0.10, our method naturally increased in the average number of peaks detected; VS cannot be tuned in this way. Conversely, for SA data the average number of peaks detected tended to be less than one for our method (with the exception of the German series).

The results are fairly similar for the quartic and TH kernels. Although the M7, M8, and VS diagnostics have slightly better performance than our spectral peak procedure with $\alpha = 0.10$, it is important to note that our method provides a level of detail that M7 and M8 cannot replicate, while the VS diagnostic does not provide a p -value for any of the peaks (neither do M7 or M8). Overall, we find the results to be very encouraging and informative.

5. Conclusion

This paper presents an innovative approach to the statistical identification of spectral peaks. The convexity diagnostic computes an average of the periodogram weighted by the second derivative of a typical kernel, such as the Tukey-Hanning lag window. Implicitly this type of statistic involves a comparison of an average of the periodogram near a given frequency to its average somewhat further out; this follows from the general shape of $\hat{A}_{\beta, \mu}$. The slope diagnostic helps to screen out cases where there is negative convexity but also a large increase/decrease in the spectrum. That the method actually works as intended is borne out by the simulations and analysis results reported in Tables 1-10.

Table 5
Data analyses for 35 Manufacturing Series (U.S. Census Bureau) comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyses – Manufacturing Series												
series	Original Data				VS	M7	M8	SA Data				VS
	H-FWER 0.10/0.05		H-FWER 0.10/0.10					H-FWER 0.10/0.05		H-FWER 0.10/0.10		
	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
M ₁	2345	1235	2345	1235	12	0.24	0.39	∅	∅	∅	∅	∅
M ₂	12345	12345	12345	12345	1235	0.20	0.32	∅	∅	∅	∅	∅
M ₃	12345	12345	12345	12345	1235	0.28	0.46	∅	∅	∅	∅	∅
M ₄	∅	∅	12345	∅	12	0.28	0.44	∅	∅	∅	∅	∅
M ₅	12345	12345	12345	12345	12345	0.27	0.47	∅	∅	12345	12345	∅
M ₆	∅	∅	123	123	12	0.28	0.49	∅	∅	∅	∅	∅
M ₇	∅	∅	123	123	24	0.50	0.79	∅	∅	∅	∅	∅
M ₈	12345	12345	12345	12345	12345	0.18	0.37	∅	∅	∅	∅	∅
M ₉	12345	12345	12345	12345	124	0.42	0.73	∅	∅	∅	∅	∅
M ₁₀	∅	∅	∅	∅	1	0.38	0.72	∅	∅	∅	∅	∅
M ₁₁	∅	∅	12345	1234	123	0.15	0.27	∅	∅	∅	∅	∅
M ₁₂	1234	1234	12345	12345	1234	0.30	0.54	∅	∅	∅	∅	∅
M ₁₄	∅	∅	1234	1234	1234	0.24	0.39	∅	∅	∅	∅	∅
M ₁₅	12345	12345	12345	12345	12345	0.23	0.43	∅	∅	∅	∅	∅
M ₁₆	1234	1234	1234	1234	1234	0.23	0.40	∅	∅	∅	∅	∅
M ₁₇	∅	∅	1234	12345	12	0.64	0.66	∅	∅	∅	∅	∅
M ₁₈	12345	12345	12345	12345	245	0.20	0.37	∅	∅	∅	∅	∅
M ₁₉	∅	∅	∅	∅	4	0.86	1.00	∅	∅	∅	∅	∅
M ₂₀	∅	∅	∅	12345	4	0.56	0.84	∅	∅	∅	∅	∅
M ₂₁	12345	12345	12345	12345	1234	0.37	0.58	∅	∅	∅	∅	∅
M ₂₂	12345	12345	12345	12345	1234	0.26	0.45	∅	∅	∅	∅	∅
M ₂₃	12345	12345	12345	12345	1234	0.20	0.47	∅	∅	∅	∅	∅
M ₂₄	12345	12345	12345	12345	2345	0.26	0.43	∅	∅	∅	∅	∅
M ₂₅	12345	12345	12345	12345	12345	0.27	0.42	∅	∅	∅	∅	∅
M ₂₆	12345	12345	12345	12345	1235	0.37	0.62	∅	∅	∅	∅	∅
M ₂₇	1345	1234	1345	1234	2345	0.25	0.22	∅	∅	∅	∅	∅
M ₂₈	∅	∅	∅	∅	24	0.57	0.44	∅	∅	∅	∅	∅
M ₂₉	∅	12345	12345	12345	24	0.78	1.13	∅	∅	∅	∅	∅
M ₃₀	123	1234	12345	12345	245	0.45	0.65	∅	∅	∅	∅	∅
M ₃₁	∅	∅	123	123	4	0.64	0.46	∅	∅	1234	1234	∅
M ₃₂	1235	12345	1235	12345	12345	0.21	0.37	∅	∅	∅	∅	∅
M ₃₃	12345	12345	12345	1234	1234	0.24	0.38	∅	∅	∅	∅	∅
M ₃₄	12345	12345	12345	12345	234	0.46	0.85	∅	∅	∅	∅	∅
M ₃₅	12345	12345	12345	12345	2345	0.25	0.66	∅	∅	∅	∅	∅
M ₃₆	12345	12345	12345	12345	123	1.32	1.56	∅	∅	∅	∅	∅
Total “correct”	23/35	24/35	31/35	31/35	35/35	34/35	33/35	35/35	35/35	33/35	33/35	35/35
Average Number	3.09	3.29	4.09	4.09	3.23			0	0	0.26	0.26	0

Table 6
Data analyses for 30 U.S. Census Bureau Series (10 Housing, 10 Import/Export and 10 Retail Sales) comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyses – Manufacturing Series												
series	Original Data						SA Data					
	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
MW1Fam	12345	12345	12345	12345	12	0.13	0.25	Ø	Ø	Ø	Ø	Ø
NW1Tot	12345	12345	12345	12345	12	0.18	0.31	Ø	Ø	Ø	Ø	Ø
NE1Fam	12345	12345	12345	12345	12	0.16	0.33	Ø	Ø	Ø	Ø	Ø
NE1Tot	12345	12345	12345	12345	123	0.25	0.27	Ø	Ø	Ø	Ø	Ø
S1Fam	12345	12345	12345	12345	125	0.22	0.47	Ø	Ø	Ø	Ø	Ø
S1Tot	124	124	1245	1245	125	0.29	0.57	Ø	Ø	Ø	Ø	Ø
US1Fam	12345	12345	12345	12345	125	0.17	0.39	Ø	Ø	Ø	Ø	Ø
US1Tot	12345	12345	12345	12345	125	0.20	0.42	Ø	Ø	Ø	Ø	Ø
W1Fam	1234	1234	12345	12345	125	0.21	0.44	Ø	Ø	Ø	Ø	Ø
W1Tot	1234	1234	12345	1234	12	0.27	0.56	Ø	Ø	Ø	Ø	Ø
Total "correct"	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Import/Export Series												
series	Original Data						SA Data					
	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
M00120	12345	Ø	12345	12345	125	0.23	0.48	Ø	Ø	Ø	Ø	Ø
M00190	12345	12345	12345	12345	1235	0.38	0.59	Ø	Ø	Ø	Ø	Ø
M3000	12345	12345	12345	12345	234	0.48	0.95	Ø	Ø	Ø	Ø	Ø
M3010	1234	1234	1234	12345	2345	0.52	0.88	Ø	Ø	Ø	Ø	Ø
M12060	12345	12345	12345	12345	123	0.53	0.77	Ø	Ø	Ø	Ø	Ø
X3	12345	12345	12345	12345	2345	0.57	0.94	Ø	Ø	Ø	Ø	Ø
X00300	134	134	134	134	2	0.56	0.97	Ø	Ø	Ø	Ø	Ø
X3020	12345	12345	12345	12345	12345	0.39	0.70	Ø	Ø	Ø	Ø	Ø
X3022	12345	12345	12345	12345	23	0.69	1.04	Ø	Ø	Ø	Ø	Ø
X10140	1234	1234	1234	1234	15	0.29	0.47	Ø	Ø	Ø	Ø	Ø
Total "correct"	10/10	9/10	10/10	10/10	10/10	10/10	9/10	10/10	10/10	10/10	10/10	10/10
Retail Series												
series	Original Data						SA Data					
	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
s0b441x0	12345	12345	12345	12345	135	0.22	0.41	Ø	Ø	Ø	Ø	Ø
s0b 44000	12345	12345	12345	12345	2345	0.12	0.26	Ø	Ø	Ø	Ø	Ø
s0b 44100	12345	12345	12345	12345	135	0.21	0.40	Ø	Ø	Ø	Ø	Ø
s0b 44130	12345	12345	12345	12345	1235	0.21	0.42	Ø	Ø	Ø	Ø	Ø
s0b 44200	12345	12345	12345	12345	12345	0.13	0.27	Ø	Ø	Ø	Ø	Ø
s0b 44300	1234	12345	1234	12345	12345	0.12	0.18	Ø	Ø	Ø	Ø	Ø
s0b 44312	1234	1234	1234	1234	12345	0.31	0.48	Ø	Ø	Ø	Ø	Ø
s0b 44400	12345	12345	12345	12345	1235	0.16	0.32	Ø	Ø	Ø	Ø	Ø
s0b 44410	12345	12345	12345	12345	1235	0.14	0.32	Ø	Ø	Ø	Ø	Ø
s0b 44500	12345	12345	12345	12345	235	0.14	0.23	Ø	Ø	Ø	Ø	Ø
Total "correct"	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Grand Total "correct"	30/30	29/30	30/30	30/30	30/30	30/30	29/30	30/30	30/30	30/30	30/30	30/30
Average Number	4.67	4.5	4.77	4.8	3.23			0	0	0	0	0

Table 7

Data analyses for 22 German OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyses – OECD DEU												
Series DEU	Original Data				SA Data						VS	
	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		
	quartic	TH	quartic	TH				quartic	TH	quartic		TH
PRMNCG03	12345	12345	12345	12345	12345	0.17	0.33	∅	∅	∅	∅	∅
PRMNCS01	1234	1234	1234	1234	235	0.25	0.79	∅	∅	∅	∅	∅
PRMNIG01	12345	12345	12345	12345	12345	0.28	0.48	∅	∅	∅	∅	∅
PRMNT001	134	134	134	134	12345	0.26	0.46	∅	∅	∅	∅	∅
PRMNVG01	1235	1235	1235	1235	2345	0.29	0.46	∅	∅	∅	∅	∅
SLMNC01	1245	1235	1245	1235	23	0.22	0.40	∅	∅	∅	∅	∅
SLMNCN01	1345	2345	1345	2345	123	0.37	0.72	∅	∅	∅	∅	∅
SLMNDM01	2345	2345	2345	2345	123	0.32	0.63	∅	∅	∅	∅	∅
SLMNEX01	12345	12345	12345	12345	12	0.32	0.51	1:5	∅	12345	12345	∅
SLMNIG01	2345	2345	2345	2345	123	0.21	0.65	∅	∅	∅	∅	∅
SLMNT001	245	345	245	345	23	0.20	0.66	∅	∅	234	∅	∅
SLRTRC01	1345	1345	1345	1345	1234	0.19	0.51	∅	∅	∅	∅	∅
SLRTO01	12345	12345	12345	12345	12345	0.12	0.25	∅	∅	∅	∅	∅
SLRTO02	12345	12345	12345	12345	12345	0.13	0.29	∅	∅	∅	∅	∅
SLWHT001	2345	2345	2345	2345	123	0.20	0.62	∅	∅	134	123	∅
SLWHT002	2345	2345	2345	2345	123	0.20	0.62	∅	∅	134	123	∅
UNLVRG01	23	23	23	23	124	0.23	0.48	∅	∅	1	∅	5
UNLVSUMA	345	345	345	345	12	0.30	0.53	12	∅	1245	12345	∅
UNLVSUTT	234	234	234	234	12	0.24	0.53	∅	∅	45	45	25
UNRTRG01	235	1245	235	1245	123	0.19	0.59	∅	∅	2345	2345	∅
XTEXVA01	1234	1234	1234	1234	23	0.28	0.77	∅	∅	∅	∅	∅
XTIMVA01	234	1234	2345	12345	23	0.31	0.95	∅	∅	∅	∅	∅
Total "correct"	22/22	22/22	22/22	22/22	22/22	22/22	22/22	20/22	22/22	14/22	16/22	20/22
Average Number	3.86	3.95	3.91	4.00	3.23			0	0	1.14	1.00	0.14

Table 8

Data analyses for 15 Euro-area OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyses – OECD EMU												
series EMU	Original Data						SA Data				VS	
	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		
	quartic	TH	quartic	TH				quartic	TH	quartic		TH
PRCNT001	345	1345	345	1345	12345	0.14	0.44	∅	∅	∅	∅	∅
PRINT001	12345	12345	12345	12345	12345	0.10	0.23	∅	∅	∅	∅	∅
PRMNCG03	1245	1245	1245	1245	12345	0.12	0.32	1	1	1234	1234	∅
PRMNCS01	1234	12345	1234	12345	1234	0.22	0.47	∅	∅	∅	∅	∅
PRMNIG01	12345	12345	12345	12345	2345	0.15	0.27	∅	∅	∅	∅	∅
PRMNT001	2345	2345	2345	2345	2345	0.14	0.23	∅	∅	1	∅	∅
PRMNVG01	1234	12345	1234	12345	2345	0.13	0.23	∅	∅	∅	∅	∅
SLMNCN02	12345	12345	12345	12345	123	0.31	0.57	∅	∅	∅	∅	∅
SLMNIG02	12345	12345	12345	12345	23	0.21	0.45	∅	∅	∅	∅	∅
SLMNT002	1345	2345	1345	2345	23	0.20	0.41	24	∅	24	34	∅
SLMNVG02	12345	12345	12345	12345	2345	0.17	0.30	1	1	1245	1345	∅
SLRTO001	12345	12345	12345	12345	12345	0.05	0.12	∅	∅	∅	∅	∅
SLRTO002	12345	12345	12345	12345	12345	0.05	0.11	∅	∅	∅	∅	∅
XTEXVA01	1345	2345	1345	2345	23	0.31	0.57	∅	∅	12	12	∅
XTIMVA01	2345	2345	2345	2345	23	0.40	0.72	∅	∅	∅	∅	∅
Total "correct"	15/15	15/15	15/15	15/15	15/15	15/15	15/15	12/15	13/15	10/15	11/15	15/15
Average Number	4.40	4.60	4.40	4.60	3.73			0.20	0.13	0.87	0.80	0

Table 9

Data analyses for 11 French OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyses – OECD FRA												
series FRA	Original Data				VS	M7	M8	SA Data				VS
	H-FWER 0.10/0.05		H-FWER 0.10/0.10					H-FWER 0.10/0.05		H-FWER 0.10/0.10		
	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
PRAFAG01	12345	12345	12345	12345	123	0.13	0.29	∅	∅	∅	∅	∅
PRNCTO01	2345	2345	2345	2345	235	0.14	0.44	∅	∅	∅	∅	∅
PRMNCG01	12345	12345	12345	12345	234	0.15	0.38	1	1	1	1	∅
PRMNCS01	12345	12345	12345	12345	234	0.25	0.58	∅	∅	∅	∅	∅
PRMNIG01	12345	12345	12345	12345	12345	0.11	0.26	∅	∅	∅	∅	∅
PRMNT001	12345	12345	12345	12345	12345	0.16	0.29	123	123	123	1234	∅
PRMNV001	12345	12345	12345	12345	12345	0.24	0.34	∅	∅	1245	1245	∅
SLRTR001	1345	2345	1345	2345	123	0.27	0.71	∅	∅	∅	∅	∅
SLRTO002	12345	12345	12345	12345	12345	0.16	0.36	∅	∅	∅	∅	∅
XTEXVA01	1345	1345	1345	1345	23	0.14	0.44	∅	∅	∅	∅	∅
XTIMVA01	1245	1245	1245	1245	23	0.18	0.54	∅	∅	∅	∅	∅
Total “correct”	11/11	11/11	11/11	11/11	11/11	11/11	11/11	9/11	9/11	8/11	8/11	11/11
Average Number	4.64	4.64	4.64	4.64	3.55			0.36	0.36	0.73	0.81	0

Table 10

Data analyses for 17 Great Britain OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyses – OECD GBR												
series GBR	Original Data				VS	M7	M8	SA Data				VS
	H-FWER 0.10/0.05		H-FWER 0.10/0.10					H-FWER 0.10/0.05		H-FWER 0.10/0.10		
	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
PPIAMP01	1234	1234	1234	1234	24	0.56	1.58	∅	∅	∅	∅	∅
PPIAMP02	1	1	123	123	2	0.53	0.92	∅	∅	∅	∅	∅
PPIPFU01	1345	12345	1345	12345	12	0.64	0.59	∅	∅	∅	∅	∅
PRINTO01	1345	1345	1345	1345	23	0.16	0.40	∅	∅	∅	∅	∅
PRMNCG02	2345	2345	2345	2345	123	0.23	0.56	∅	∅	∅	∅	∅
PRMNCG03	12345	12345	12345	12345	123	0.20	0.49	∅	∅	∅	∅	∅
PRMNCS01	123	12	123	1234	12	0.68	1.31	∅	∅	∅	∅	∅
PRMNIG01	2345	2345	2345	2345	123	0.15	0.47	1	∅	∅	∅	∅
PRMNT001	1345	1345	1345	1345	23	0.17	0.45	∅	∅	∅	∅	∅
PRMNV002	12345	12345	12345	12345	12345	0.25	0.76	∅	∅	∅	∅	∅
PRMNV003	1234	1234	1234	1234	234	0.29	0.91	∅	∅	∅	∅	∅
PRMNVG01	124	134	124	134	234	0.18	0.58	∅	∅	∅	∅	∅
SLRTR003	1345	1345	1345	1345	124	0.42	0.74	124	123	124	123	∅
SLRTO002	12345	12345	12345	12345	12345	0.05	0.15	1234	∅	1234	12345	∅
UNLVRG01	12345	12345	12345	12345	1245	0.63	0.61	∅	∅	∅	∅	∅
XTEXVA01	134	134	1345	1345	23	0.34	1.02	∅	∅	∅	∅	∅
XTIMVA01	23	23	23	23	23	0.31	0.90	∅	∅	∅	∅	∅
Total “correct”	17/17	17/17	17/17	17/17	17/17	17/17	14/17	14/17	16/17	13/17	15/17	17/17
Average Number	3.76	3.76	3.94	4.06	2.76			0.47	0.18	0.53	0.47	0

For the multiple peak-testing scenario, we employ known results from the multiple testing literature (*i.e.*, the applications to controlling FWER) to combine the p -values from the five seasonal frequencies in such a way to dramatically increase statistical power, as demonstrated in Table 4. Although there is some departure in the size (Table 3) for the multiple peak testing, the results are still quite usable. On a typical batch of seasonal series, the number of Type I errors are as expected, and the power is quite decent (Tables 5-10). While our method compares quite favorably to the VS, M7, and M8 diagnostics, neither of the diagnostics provide a p -value and only the former can distinguish which spectral peaks are contributing to seasonal behavior. This aspect is important to the seasonal adjuster, who wants to know not only that there may be residual seasonality, but also at what seasonal frequencies, so as to take appropriate action to alter the seasonal adjustment filters (this can be done by smoothing over additional years, which is accomplished by changing the seasonal filters in X-11-ARIMA; alternatively, one might consider shortening the series. For current research on a model-based approach to designing SA filters targeted for specific seasonal frequencies, see Aston, Findley, McElroy, Wills, and Martin (2007)).

The choice of kernel surely has some impact on the results, although we found little difference in practice between the quartic and TH kernels; the TH may be marginally more powerful. Of course, plenty of other popular kernels may also be utilized by a practitioner, and we have only chosen two that seemed intuitive and straightforward to implement. The choice of the location μ is clearly dictated by the characterization of seasonality. Since statistical power generally decreased with β , we always recommend taking the maximal β such that the kernel supports are disjoint, which guarantees the asymptotic independence property of the various diagnostics that is crucial to our multiple testing method.

Finally, the asymptotic results require that the data be differenced to stationarity. Recognizing that economic time series are typically nonstationary, it is desirable to trend-difference seasonally adjusted data before applying our diagnostic. This differencing may dampen the detection of the first seasonal peak, so practitioners may “re-color” the data as described in Section 3.4.

Acknowledgements

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. Holan’s

research was supported by an ASA/NSF/BLS research fellowship.

Appendix

Here we derive asymptotic formulas for the statistical measures ψ_A of slope and convexity. These results can then be applied in the testing paradigm to get asymptotic critical values. Some mild conditions on the data are required for the asymptotic theory; we follow the material in Taniguchi and Kakizawa (2000, Section 3.1.1). Condition (B), due to Brillinger (1981), states that the process is strictly stationary and condition (B1) of Taniguchi and Kakizawa (2000, page 55) holds. Condition (HT), due to Hosoya and Taniguchi (1982), states that the process has a linear representation, and conditions (H1) through (H6) of Taniguchi and Kakizawa (2000, pages 55-56) hold. Assumption 1 (8) of Chiu (1988) is a summability condition on various higher order cumulants, which is satisfied, for example, by a Gaussian process with spectral density in C^2 . None of these conditions are stringent; for example, a causal linear process with fourth moments satisfies (HT). The main result is a joint convergence of any two measures $\psi_A(I)$; *e.g.*, these can be a slope and convexity measure with the same kernel A . We present the general theorem that covers these two cases.

Theorem 1 Suppose that the fourth order cumulants of $\{X_t\}$ vanish; that either condition (B) or (HT) holds; and that Assumption 1 (8) of Chiu (1988) holds. Let the kernels A and B satisfy conditions (i) through (iv) of Section 2.1. Then

$$\left\{ \begin{array}{l} \sqrt{n} \frac{(\theta_A(I) - \theta_A(f))}{\sqrt{\theta_{A^2}(I^2)/2}}, \\ \sqrt{n} \frac{(\theta_B(I) - \theta_B(f))}{\sqrt{\theta_{B^2}(I^2)/2}} \end{array} \right\} \xrightarrow{L} N(0, V)$$

as $n \rightarrow \infty$. Here 0 denotes the zero vector $(0, 0)'$, and V is a 2×2 matrix with entries

$$V_{11} = V_{22} = 1 \quad V_{12} = V_{21} = \frac{\theta_{AB}(f^2)}{\sqrt{\theta_{A^2}(f^2)\theta_{B^2}(f^2)}}.$$

Proof. First we establish that $\theta_{A^2}(I^2) \xrightarrow{a.s.} 2\theta_{A^2}(f^2)$. Since the kernel A is continuous in an interval (such as $[\mu - \beta/2, \mu + \beta/2]$), this result follows directly from Corollary 1 of Chiu (1988), noting that they deal with the Riemann sums approximation to the integral functional (Chiu (1988) also defines the periodogram with a 2π factor). Of course the same results holds with B in place of A . Secondly, consider the joint convergence of $\theta_A(I)$ and

$\theta_B(I)$. We use the Cramer-Wold device, and apply Lemma 3.1.1 of Taniguchi and Kakizawa (2000), appropriately generalized to include non-even functions (cf. Theorem 3 of Chiu (1988)). Hence for any x, y real,

$$\sqrt{n} \left(x \frac{(\theta_A(I) - \theta_A(f))}{\sqrt{\theta_{A^2}(f^2)}} + y \frac{(\theta_B(I) - \theta_B(f))}{\sqrt{\theta_{B^2}(f^2)}} \right) \xrightarrow{L} N \left(0, \frac{1}{2\pi} \int_{-\pi}^{\pi} (C(\lambda)C(-\lambda) + C^2(\lambda)) f^2(\lambda) d\lambda \right)$$

using Slutsky's Theorem (Bickel and Doksum 1977), where the kernel C is defined by

$$C(\lambda) = \frac{x}{\sqrt{\theta_{A^2}(f^2)}} A(\lambda) + \frac{y}{\sqrt{\theta_{B^2}(f^2)}} B(\lambda).$$

Clearly $C(\lambda)C(-\lambda) = 0$ and

$$\begin{aligned} C^2(\lambda) &= \frac{x^2}{\theta_{A^2}(f^2)} A^2(\lambda) \\ &+ 2 \frac{xy}{\sqrt{\theta_{A^2}(f^2)} \sqrt{\theta_{B^2}(f^2)}} A(\lambda) B(\lambda) \\ &+ \frac{y^2}{\theta_{B^2}(f^2)} B^2(\lambda). \end{aligned}$$

By taking x and y to be zero and one in various combinations, we deduce the stated variance matrix V .

Next we discuss the multiple-peak testing scenario. So suppose that we have a finite collection of kernels A_i for $i = 1, 2, \dots, d$, each of which satisfies the assumptions of Section 2. Then we can easily generalize Theorem 1 from two to d kernels as follows. The asymptotic covariance matrix V will have ij^{th} entry

$$\frac{\theta_{A_i A_j}(f^2)}{\sqrt{\theta_{A_i^2}(f^2) \theta_{A_j^2}(f^2)}}.$$

Thus, if the support for any two kernels is disjoint we obtain asymptotic independence, and can therefore invoke the H-FWER multiple testing procedure.

References

- Aston, J., Findley, D., McElroy, T., Wills, K. and Martin, D. (2007). New ARIMA Models for Seasonal Time Series and Their Application to Seasonal Adjustment and Forecasting. *SRD Research Report No. RRS 2007-14*, U.S. Census Bureau.
- Bell, W., and Hillmer, S. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-320.
- Bickel, P., and Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Englewood Cliffs, New Jersey: Prentice Hall.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.
- Brillinger, D. (1981). *Time Series Data Analysis and Theory*. San Francisco: Holden-Day.
- Chiu, S. (1988). Weighted least squares estimators on the frequency domain for the parameters of a time series. *The Annals of Statistics*, 16, 1315-1326.
- Evans, T., Holan, S. and McElroy, T. (2006). Evaluating measures for assessing spectral peaks. *2006 Proceedings American Statistical Association*, [CD-ROM]: Alexandria, VA.
- Findley, D. (2006). Personal communication.
- Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. and Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177 (with discussion).
- Hosoya, Y., and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics*, 10, 132-153.
- Lothian, J., and Morry, M. (1978). A test for identifiable seasonality when using the X-11-ARIMA program. Working Paper, Time Series Research and Analysis Division, Statistics Canada.
- Maravall, A., and Caporello, G. (2004). Program TSW: Revised Reference Manual. *Working Paper 2004, Research Department, Bank of Spain*. <http://www.bde.es>.
- Nerlove, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica*, 32, 241-286.
- Newton, H., and Pagano, M. (1983). A method for determining periods in time series. *Journal of the American Statistical Association*, 78, 152-157.
- Parzen, E. (1983). Autoregressive spectral estimation. *Handbook of Statistics III*, (Ed. D. Brillinger and P. Krishnaiah), Amsterdam: North Holland, 221-247.
- Priestley, M. (1981). *Spectral Analysis and Time Series*. London: Academic Press.
- Soukup, R.J., and Findley, D.F. (1999). On the spectrum diagnostics used by X-12-ARIMA to indicate the presence of trading day effects after modeling or adjustment. Also www.census.gov/pub/ts/papers/r9903s.pdf. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 144-149.
- Taniguchi, M., and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. New York City, New York: Springer-Verlag.
- U.S. Census Bureau (2002). X-12 ARIMA Reference Manual (Version 0.2.10), Washington, DC.

On the definition and interpretation of interviewer variability for a complex sampling design

Siegfried Gabler and Partha Lahiri¹

Abstract

Interviewer variability is a major component of variability of survey statistics. Different strategies related to question formatting, question phrasing, interviewer training, interviewer workload, interviewer experience and interviewer assignment are employed in an effort to reduce interviewer variability. The traditional formula for measuring interviewer variability, commonly referred to as the interviewer effect, is given by $ieff := deff_{int} = 1 + (\bar{n}_{int} - 1)\rho_{int}$, where ρ_{int} and \bar{n}_{int} are the intra-interviewer correlation and the simple average of the interviewer workloads, respectively. In this article, we provide a model-assisted justification of this well-known formula for equal probability of selection methods (epsem) with no spatial clustering in the sample and equal interviewer workload. However, spatial clustering and unequal weighting are both very common in large scale surveys. In the context of a complex sampling design, we obtain an appropriate formula for the interviewer variability that takes into consideration unequal probability of selection and spatial clustering. Our formula provides a more accurate assessment of interviewer effects and thus is helpful in allocating more reasonable amount of funds to control the interviewer variability. We also propose a decomposition of the overall effect into effects due to weighting, spatial clustering and interviewers. Such a decomposition is helpful in understanding ways to reduce total variance by different means.

Key Words: Interviewer effect; Interviewer workloads; Intra-interviewer correlation; Spatial clustering; Unequal weighting.

1. Introduction

A major source of measurement errors in surveys is due to the interviewer. This fact was recognized as early as 1929 by Rice and later by many survey researchers. Factors such as the quality of questionnaire design and the interviewer can influence the interviewer effects on survey statistics.

The interviewer can introduce homogeneity in survey data, which generally reduces the effective sample size and thereby increases the total variance of a survey estimator. The within interviewer homogeneity has been traditionally measured by the intra-interviewer correlation coefficient ρ_{int} . The magnitude of the intra-interviewer correlation was studied by many researchers, mostly in the context of telephone surveys without any spatial clustering effects (Kish 1962; Gray 1956; Hanson and Marks 1958; Tucker 1983; Groves and Magilavy 1986; Heeb and Gmel 2001, and others). Researchers have argued that the nature of the survey items may affect the value of ρ_{int} . Attitude items and complex factual items are considered more sensitive to the intra-interviewer correlation than simple factual items are (Collins and Butcher 1982; Feather 1973; Fellegi 1964; Gray 1956; Hansen, Hurwitz and Bershad 1961). According to Groves (1989), values above 0.1 are seldom observed. See Schnell and Kreuter (2005) for further discussion on this issue.

As noted by several researchers, the standard interviewer effect formula $1 + (\bar{n}_{int} - 1)\rho_{int}$ suggests that even with a small intra-interviewer correlation, the interviewer effect could be substantial simply due to a high average interviewer workload. For example, when $\rho_{int} = 0.01$ and $\bar{n}_{int} = 70$ we have $ieff = 1.69$ (Schnell and Kreuter 2005). Note that a high average interviewer workload (e.g., between 60 and 70) is very common in telephone surveys (Tucker 1983; Groves and Magilavy 1986). For the European Social Survey, Philippens and Loosveldt (2004) provided box plots of the intra-interviewer correlations and the interviewer workloads for 18 participating countries.

The interviewer effect or variance is generally defined as the inflation to the total variance caused solely by the interviewers. For an epsem design with equal interviewer workload, the interviewer variance for the sample mean is simply given by $1 + (n_{int} - 1)\rho_{int}$, where n_{int} is the common interviewer workload. For complex surveys with unequal interviewer workload, survey researchers frequently use a simple modification of this formula where the common interviewer workload is replaced by the average interviewer workload, i.e., the formula $1 + (\bar{n}_{int} - 1)\rho_{int}$. In Section 2, we argue that this standard formula $1 + (\bar{n}_{int} - 1)\rho_{int}$ cannot be interpreted as an inflation to the total variance caused by the interviewers even for an epsem design with unequal interviewer workload. In Sections 2-4, we observe that the interviewer variance definition depends

1. Siegfried Gabler, GESIS, P.O. Box 12 21 55, 68072 Mannheim, Germany. E-mail: siegfried.gabler@gesis.org; Partha Lahiri, University of Maryland, College Park, U.S.A. E-mail: plahiri@survey.umd.edu.

on the nature of the complex sampling design and also on the interviewer workload assignment. In this paper, we provide appropriate definitions of the interviewer variance in different survey scenarios. A reliable definition of the interviewer variance is helpful in determining actions that need to be taken in order to reduce interviewer variability. This paper is foremost applicable to the planning of surveys rather than analyzing survey data. In other words, in this paper we have concentrated on the definitions and interpretation of the interviewer variability and not on estimating it from a given survey.

In Section 2, we consider an epsem design with no spatial clustering and provide a model-assisted interpretation of $ieff$. We show that for the equal interviewer workload $ieff$ is simply the ratio of the variances of the sample mean under a correlated model that accounts for the homogeneity of the observations collected by the same interviewer and a simple uncorrelated model that fails to account for such homogeneity. Thus, multiplying the variance of the sample mean for simple random sampling by the $ieff$ one can obtain the total variance of the sample mean that incorporates both the sampling and the interviewer variability. This is a very intuitive interpretation of $ieff$ and complements the model-assisted justification given earlier by Kish (1962). In this section, we also show that for an epsem design $ieff$ is lower than the model-assisted interviewer effect formula if the interviewer workload varies and the intra-interviewer correlation is positive. Thus, the survey designer who uses $ieff$ would give less effort to control interviewer variability than is really needed. In this situation, an appropriate interviewer effect formula can be obtained from $ieff$ when a weighted average interviewer workload is used in place of the usual simple average.

In Section 3, we entertain the possibility of unequal weighting but no spatial clustering. We obtain a model-assisted interpretation for $ieff$ if and only if the respondents interviewed by the same interviewer share the same sampling weight and the interviewer workload is inversely proportional to the square of the common weight for the interviewer. Interestingly, unlike the epsem design, equal interviewer workload does not necessarily guarantee a model-assisted interpretation for $ieff$. When there is an equal interviewer workload and there is at least one interviewer for which the respondents do not all share the same sampling weight, we show that $ieff$ is always higher than the model-assisted formula. We also point out the factors that cause the difference between these two formulae. These results have a practical relevance in terms of saving survey costs. To be specific, the survey designer who uses $ieff$ is likely to allocate more funds to control interviewer variability than is really needed. We have also

cited some situations where $ieff$ could have an underestimation problem and thus survey designers who use $ieff$ could give less emphasis to control the interviewer effects. Our formula provides a more accurate assessment of interviewer variability and thus is helpful in the allocation of more reasonable amount of funds to control the interviewer variability. Furthermore, the change in planning formulae will affect the sample size.

In many large scale sample surveys, due to various organizational and financial reasons such as the absence of a general population register or to reduce the overall survey costs, a multi-stage clustered sampling design is considered to be a cost-efficient alternative to simple random sampling. Under a multi-stage clustered sampling design, respondents who live in close spatial proximity of each other get selected. Respondents living in the same spatial cluster tend to share similar attitudes because of their similar socio-economic background and hence increase the internal homogeneity of the survey data. This spatial homogeneity violates the iid (independently identically distributed) assumption frequently used in standard statistical inferential procedures and so does the clustering within the interviewers. This fact has been recognized by many survey researchers and adjustments to various statistical procedures and the related software issues have been addressed in the literature (see Rao and Scott 1984; Skinner, Holt and Smith 1989; Biemer and Trewin 1997; Chambers and Skinner 2003; among others). In Section 4, we present a new definition of the interviewer variability in the presence of unequal weighting and spatial clustering. In the presence of spatial clustering, we argue that $ieff$ generally has a tendency to overestimate the interviewer variability. Thus for complex surveys involving spatial clustering, $ieff$ may unnecessarily give a false alarm regarding the magnitude of the interviewer variability.

In Section 5, we discuss the effects due to the combined effects of weighting, spatial clustering and the interviewer. The formula for overall effects offers an accurate determination of the sample size at the planning stage. We provide a nice factorization of the overall effects into the effects due to weighting, clustering and interviewer. Such a decomposition of the overall effects can be useful in understanding ways to reduce the total variance by different means. In discussing Verma, Scott and O'Muircheartaigh (1980), Hedges mentioned the need for such an overall effect formula. We generalize a formula earlier proposed by Davis and Scott (1995) to a non-epsem design and for a general correlation model valid for both discrete and continuous data. We present proofs of all the technical results in the Appendix.

2. EPSEM design with no spatial clustering

Let y_{ik} denote the observation obtained from the k^{th} respondent interviewed by the i^{th} interviewer ($i = 1, \dots, I$; $k = 1, \dots, n_i$). Define $n = \sum_{i=1}^I n_i$, the total sample size, $\bar{y} = 1/n \sum_{i=1}^I \sum_{k=1}^{n_i} y_{ik}$, the unweighted sample mean, and $\bar{n}_{\text{int}}(\mathbf{a}) = \sum_{i=1}^I a_i n_i$, a weighted average of the interviewer workload, where a_i is an arbitrary weight attached to the i^{th} interviewer workload and $\mathbf{a} = (a_1, \dots, a_I)$.

We shall first provide a model-assisted justification of the traditional interviewer effect formula, *i.e.*, $ieff = 1 + (\bar{n}_{\text{int}} - 1)\rho_{\text{int}}$, where \bar{n}_{int} is the unweighted average of interviewer workload. Note that $\bar{n}_{\text{int}} = \bar{n}_{\text{int}}(\mathbf{a}_0)$, with $\mathbf{a}_0 = (a_{01}, \dots, a_{0I})$, $a_{0i} = 1/I$ and $ieff = ieff(\mathbf{a}_0)$. Using Result 1 given in the Appendix, we get

$$ieff(\mathbf{a}_1) = \frac{\text{Var}_{M_2}(\bar{y})}{\text{Var}_{M_1}(\bar{y})} = 1 + [\bar{n}_{\text{int}}(\mathbf{a}_1) - 1]\rho_{\text{int}},$$

where $\mathbf{a}_1 = (a_{11}, \dots, a_{1I})$, with $a_{1i} = n_i/n$. In the above, $\text{Var}_{M_1}(\bar{y})$ and $\text{Var}_{M_2}(\bar{y})$ are the variances of \bar{y} under the following two models, respectively,

$$M_1: \text{Cov}(y_{ik}, y_{i'k'}) = \begin{cases} \sigma^2 & \text{if } i = i', k = k', \\ 0 & \text{otherwise,} \end{cases}$$

$$M_2: \text{Cov}(y_{ik}, y_{i'k'}) = \begin{cases} \sigma^2 & \text{if } i = i', k = k', \\ \rho_{\text{int}}\sigma^2 & \text{if } i = i', k \neq k', \\ 0 & \text{otherwise.} \end{cases}$$

Note that unlike model M_1 , model M_2 introduces homogeneity of the observations collected by the same interviewer.

Remark 2.1: It follows from the corollary to Result 1, given in the Appendix, that for $\rho_{\text{int}} > 0$, $ieff(\mathbf{a}_1) = ieff$ if and only if $n_i = n/I$ for all i , *i.e.*, if and only if each interviewer has the same workload. For the balanced case, Kish (1962) provided a model-assisted justification of $ieff$ using a linear mixed model, which is a special case of M_2 . For the unbalanced case, it is interesting to note the similarity between the interviewer variability formula $ieff(\mathbf{a}_1)$ and the design effects formula given in (A3) of Holt in discussing Verma *et al.* (1980).

Remark 2.2: It follows from the corollary to Result 1 that if $\rho_{\text{int}} > 0$ and n_i 's are not equal then $ieff(\mathbf{a}_1) > ieff$.

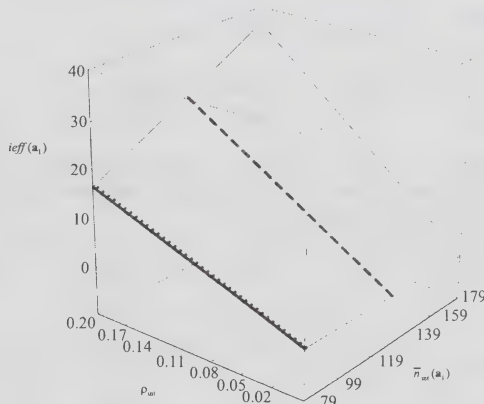
In the following example, we demonstrate the extent to which $ieff(\mathbf{a}_1)$ and $ieff$ could differ for different interviewer workload patterns.

Example 1: In Table 1, we consider three different workload assignments for ten interviewers, each with $n = 790$. Case A) represents the most variable workload assignment with a standard deviation = 68.3; Case B) is nearly balanced with a standard deviation = 9.5; Case C) corresponds to the equal interviewer assignment.

Table 1
Three different interviewer workload assignments (Example 1)

Interviewer	Interviewer workload pattern		
	A)	B)	C)
1	4	70	79
2	10	70	79
3	20	70	79
4	34	70	79
5	52	70	79
6	74	88	79
7	100	88	79
8	130	88	79
9	164	88	79
10	202	88	79
n	790	790	790
$\bar{n}_{\text{int}}(\mathbf{a}_1)$	132	80	79

Let $ieff(\mathbf{a}_{1,A})$, $ieff(\mathbf{a}_{1,B})$, and $ieff(\mathbf{a}_{1,C}) = ieff$ denote $ieff(\mathbf{a}_1)$, the model-assisted interviewer variance formula corresponding to the cases A, B and C, respectively. For $\rho_{\text{int}} > 0$ the function $ieff(\mathbf{a}_1)$ is Schur-convex, which explains the fact $ieff(\mathbf{a}_{1,A}) \geq ieff(\mathbf{a}_{1,B}) \geq ieff(\mathbf{a}_{1,C}) = ieff$. Figure 1 provides the values of the interviewer variance obtained from the standard formula (*i.e.*, $ieff$) and our model-assisted interview variance formula for all combinations of the two influencing factors, *i.e.*, weighted average of interviewer workload and the intra-interviewer correlation. From Figure 1, it is interesting to note that $ieff$ could underreport by about 100%.



Interviewer A: Dashes,
Interviewer B: Spaced dots and
Interviewer C: Solid line

Figure 1 A graph of $ieff(\mathbf{a}_1)$ vs. ρ_{int} for different $\bar{n}_{\text{int}}(\mathbf{a}_1)$

3. Unequal weighting with no spatial clustering

In this section, we consider the situation when we have unequal weights. Let w_{ik} be the survey weight attached to the k^{th} respondent interviewed by the i^{th} interviewer. In this situation, a weighted mean $\bar{y}_w = \sum_i \sum_k w_{ik} y_{ik} / \sum_i \sum_k w_{ik}$ is a popular estimator of the finite population mean (See Brewer 1963; Hájek 1971) and the model-assisted interviewer variance formula is given by

$$ieff_w = \frac{\text{Var}_{M_2}(\bar{y}_w)}{\text{Var}_{M_1}(\bar{y}_w)} = 1 + \rho_{\text{int}} \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right).$$

See Result 1 given in the Appendix.

Define $\bar{w}_i = 1/n_i \sum_{k=1}^{n_i} w_{ik}$, the average survey weight for the i^{th} interviewer and $\sigma_i^2 = 1/n_i \sum_k w_{ik}^2 - \bar{w}_i^2$, the variance of the survey weights for the i^{th} interviewer. It can be shown that

$$ieff_w = 1 + \rho_{\text{int}} (\bar{n}_w - 1),$$

where

$$\bar{n}_w = \frac{\sum_i n_i^2 \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2 + \sum_i n_i \sigma_i^2}.$$

Note that, in general, $ieff_w$ cannot be written in the form $ieff_w = 1 + \rho_{\text{int}} (\bar{n}_{\text{int}}(\mathbf{a}) - 1)$ with $\sum_i a_i = 1$.

Remark 3.1: From Result 2 in the Appendix, we have

$$ieff_w \leq ieff(\mathbf{a}_2),$$

where

$$\mathbf{a}_2 = (a_{21}, \dots, a_{2I}), \text{ with } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2}.$$

In the above, for $\rho_{\text{int}} > 0$, $ieff_w = ieff(\mathbf{a}_2)$ if and only if all σ_i^2 are zero. Thus, $ieff(\mathbf{a}_2)$ can be interpreted as a conservative interviewer variance.

Equality holds if and only if $w_{ik} = \bar{w}_i$ for all i and k in which case

$$ieff_w = ieff(\mathbf{a}_2^*),$$

where

$$\mathbf{a}_2^* = (a_{21}^*, \dots, a_{2I}^*), \text{ with } a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}.$$

Thus, the formulae $ieff_w$ and $ieff(\mathbf{a}_2^*)$ are equivalent if and only if the survey weights are all the same for a given

interviewer. One example of such a design is an epsem design for which we have

$$a_{2i}^* = \frac{n_i}{n}$$

and

$$ieff_w = ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_1).$$

Now we shall try to understand the factors that explain the difference between $ieff_w$ and $ieff$. To this end, define

$\bar{w} = 1/n \sum_{i=1}^I \sum_{k=1}^{n_i} w_{ik} = \sum_{i=1}^I n_i/n \bar{w}_i$, the average survey weight for all interviewers,

$SSB = \sum_{i=1}^I n_i (\bar{w}_i - \bar{w})^2$, the between interviewer sum of squares of the survey weights,

$SSW = \sum_{i=1}^I \sum_{k=1}^{n_i} (w_{ik} - \bar{w}_i)^2 = \sum_{i=1}^I n_i \sigma_i^2$, the within interviewer sum of squares of the survey weights,

$SST = SSB + SSW$, the total sum of squares of the survey weights,

$\tau_w = SSW/SST$, an indicator of the relative contribution of the within interviewer variability of survey weights to the total variability,

$CV_w = \sqrt{SST/n} / \bar{w}$, the coefficient of variation of the survey weights in the entire sample.

It can be shown that (see Result 4)

$$ieff_w - ieff = \frac{\bar{n}_{\text{int}}}{SST + n \bar{w}^2} \left[\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SSW \right] \rho_{\text{int}} \quad (1)$$

$$= \frac{\bar{n}_{\text{int}}}{(1 + CV_w^{-2}) SST} \left[\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SSW \right] \rho_{\text{int}} \quad (2)$$

$$= \frac{\bar{n}_{\text{int}} \tau_w}{1 + CV_w^{-2}} \left(\frac{\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2}{SSW} - 1 \right) \rho_{\text{int}}. \quad (3)$$

Remark 3.2: We can use formula (1) in any situation. For epsem designs, we have

$$ieff_w - ieff = \rho_{\text{int}} \frac{\bar{n}_{\text{int}}}{n} \sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i.$$

Note that an application of the Cauchy-Schwarz inequality suggests $ieff_w - ieff \geq 0$ with equality if and only if $n_i = n/I$ for all i .

Remark 3.3: We can use (2) if $SST \neq 0$, i.e., if the design is not epsem. If $\rho_{\text{int}} > 0$, (2) implies

$$ieff_w - ieff \leq 0 \text{ if and only if } \sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{int}} - 1 \right) n_i \bar{w}_i^2 \leq SSW.$$

If high interviewer workload tends to be associated with small average survey weights and vice versa and $SSW \neq 0$, we can expect $ieff$ to be a conservative value of the actual interviewer variance $ieff_w$. In Example 2, c) and d), we have such a situation.

Now, we have $ieff_w = ieff$ if and only if $w_{ik} = \bar{w}_i$ (or, equivalently, $SSW = 0$) and $n_i \bar{w}_i^2 / \sum_i n_i \bar{w}_i^2 = 1/I$ for all i and k , i.e., $ieff_w = ieff$ if and only if $w_{ik} = \bar{w}_i$ and $\bar{w}_i \propto 1/\sqrt{n_i}$ for all i and k .

Thus, for a non-epsem design, equal interviewer workload does not necessarily provide us a model-assisted interpretation for $ieff$. For example, if the survey weights vary within at least one interviewer, we will not have a model-assisted interpretation of $ieff$. Obviously, for an epsem design the two formulae are equivalent if and only if we have equal interviewer workload.

Remark 3.4: If the interviewer workload is the same for all interviewers, we have

$$ieff_w - ieff = - \frac{\bar{n}_{int} \tau_w}{1 + CV_w^{-2}} \rho_{int}$$

(assume $SST \neq 0$). Thus, $ieff$ is a conservative value of the actual interviewer effect $ieff_w$. Furthermore, $|ieff_w - ieff|$ is an increasing function of the common interviewer workload \bar{n}_{int} and $\tau_w / (1 + CV_w^{-2})$ (for fixed CV_w^{-2} , the latter is an increasing function of τ_w). The same interviewer workload is given in Example 2 a).

Remark 3.5: We can use formula (3) if $SSW > 0$, i.e., if there is at least one interviewer for which weights are not all equal.

Example 2.

Table 2 presents eight different combinations of $(n_i, \bar{w}_i, \sigma_i^2)$. The first combination assumes equal n_i values but unequal weights. The second combination assumes $\bar{w}_i^2 \propto \sigma_i^2$. The other six combinations show all possible ordering of $\bar{n}_{int}, \bar{n}_{int}(a_1), \bar{n}_w, \bar{n}_{int}(a_2)$ and, therefore, $ieff, ieff(a_1), ieff_w, ieff(a_2)$ taking into consideration that $ieff \leq ieff(a_1)$ and $ieff_w \leq ieff(a_2)$.

Table 2
Ordering of interviewer effects formulae for several parameter combinations (Example 2); in the last column $\rho_{int} = 0.01$

	n_i	\bar{w}_i	σ_i^2	\bar{n}_{int}	$\bar{n}_{int}(a_1)$	\bar{n}_w	$\bar{n}_{int}(a_2)$	Interviewer effects	$ieff / ieff_w$
a)	25	1.022	0.299	25	25	19.20	25	$ieff = ieff(a_1) = ieff(a_2) > ieff_w$	1.003
	25	1.036	0.375						
	25	0.998	0.276						
	25	0.945	0.260						
b)	10	1	1	25	30	15	30	$ieff_w < ieff < ieff(a_1) = ieff(a_2)$	1.007
	20	1	1						
	30	1	1						
	40	1	1						
c)	10	1	1	25	30	7.5	32.5	$ieff_w < ieff < ieff(a_1) < ieff(a_2)$	1.023
	20	1	2						
	30	1	3						
	40	1	4						
d)	10	1	4	25	30	10	26.7	$ieff_w < ieff < ieff(a_2) < ieff(a_1)$	1.015
	20	1	3						
	30	1	2						
	40	1	1						
e)	10	4	144	25	30	1.80	11.71	$ieff_w < ieff(a_2) < ieff < ieff(a_1)$	0.998
	20	2	9						
	30	0.333	0.555						
	40	0.250	0.125						
f)	10	0.333	0.025	25	30	31.82	35.26	$ieff < ieff(a_1) < ieff_w < ieff(a_2)$	1.015
	20	0.666	0.075						
	30	1	0.125						
	40	1.333	0.175						
g)	10	1	0.010	25	30	29.13	30.10	$ieff < ieff_w < ieff(a_1) < ieff(a_2)$	0.999
	20	1	0.020						
	30	1	0.030						
	40	1	0.040						
h)	10	1	0.004	25	30	29.94	29.99	$ieff < ieff_w < ieff(a_2) < ieff(a_1)$	0.998
	20	1	0.003						
	30	1	0.002						
	40	1	0.001						

In the example, $\sum_i n_i \bar{w}_i = n$. We now explain the eight different patterns.

- a) Since all n_i are equal, $ieff = ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$. Moreover, $ieff_w$ is smaller than the rest because of the fact that $\sigma_i^2 > 0$.
- b) Since σ_i^2 are relatively large, $ieff_w < ieff$. Also, $\sigma_i^2 = c \cdot \bar{w}_i^2$ implies $ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$.
- c) Since σ_i^2 are relatively large, $ieff_w < ieff$. Moreover, since $\bar{w}_i^2 + \sigma_i^2$ and n_i are both increasing, we have $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$.
- d) Since σ_i^2 are relatively large, $ieff_w < ieff$. Since $\bar{w}_i^2 + \sigma_i^2$ is decreasing and n_i is increasing, we have $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.
- e) Since σ_i^2 are relatively large, $ieff_w < ieff$. Also, \bar{w}_i^2 and σ_i^2 are decreasing and n_i is increasing implying $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.
- f) The fact that \bar{w}_i^2 and n_i are increasing implies that $ieff_w > ieff$; since σ_i^2 and n_i are both increasing, we have $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$.
- g) Since \bar{w}_i^2 and n_i are increasing, we have $ieff_w > ieff$ and since σ_i^2 is increasing, we have $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$. Moreover, $ieff_w < ieff(\mathbf{a}_1)$ since σ_i^2 is smaller than that in f).
- h) Since \bar{w}_i^2 and n_i are increasing, we have $ieff_w > ieff$ and since σ_i^2 is decreasing, we have $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

4. Unequal weighting and spatial clustering

In this section, we obtain an appropriate interviewer variance formula in the presence of spatial clustering and unequal probability of selection. Consider the situation when more than one interviewer work independently in the same psu and the respondents in each psu are randomly assigned to the interviewers. We shall assume that no interviewer works in more than one psu. Such a design was considered in Biemer and Stokes (1985). Now we shall separate the interviewer effect from psu effect (i.e., spatial clustering) and unequal weighting. Let y_{pik} and w_{pik} be the observation and the associated survey weight for the k^{th} respondent in the p^{th} psu interviewed by the i^{th} interviewer ($p = 1, \dots, P$; $i = 1, \dots, I_p$; $k = 1, \dots, n_{pi}$). Let $n_p = \sum_{i=1}^{I_p} n_{pi}$ be the number of sampling units in psu p .

In this case, we use the following weighted average to estimate the finite population mean:

$$\bar{y}_w = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} y_{pik}}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}}.$$

Define

$$ieff_{s,w} = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_3}(\bar{y}_w)},$$

where the suffixes s and w signify the presence of spatial clustering and unequal weighting. In the above, $\text{Var}_{M_3}(\bar{y}_w)$ and $\text{Var}_{M_4}(\bar{y}_w)$ are the variances of \bar{y}_w under the following two models respectively

$$M_3: \text{Cov}(y_{pik}, y_{p'ik'}) = \begin{cases} \sigma^2 & \text{if } p=p', i=i', k=k' \\ \rho_C \sigma^2 & \text{if } p=p', k \neq k' \\ 0 & \text{otherwise} \end{cases}$$

$$M_4: \text{Cov}(y_{pik}, y_{p'ik'}) = \begin{cases} \sigma^2 & \text{if } p=p', i=i', k=k' \\ \rho_C \sigma^2 & \text{if } p=p', i \neq i' \\ \rho \sigma^2 & \text{if } p=p', i=i', k \neq k' \\ 0 & \text{if } p \neq p' \end{cases}$$

In the above, ρ_C is the intra-psu correlation and ρ is the combined interviewer and psu intra-class correlation. Define $\rho_{\text{int}} = \rho - \rho_C$, intra-interviewer correlation. Usually, $\rho_{\text{int}} > 0$.

From Result 5, we have

$$ieff_{s,w} = 1 + \rho_{\text{int}} \frac{\bar{n}_{\text{int}}(\mathbf{A}_w) - 1}{1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}_w) - 1)},$$

where

$$\mathbf{A}_w = ((a_{wpi}))_{i=1, \dots, I_p, p=1, \dots, P} \text{ and } a_{wpi} = \frac{n_{pi} \bar{w}_{pi}^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}$$

with

$$\bar{w}_{pi} = \frac{1}{n_{pi}} \sum_k w_{pik},$$

$$\bar{n}_{\text{int}}(\mathbf{A}_w) = \sum_{p=1}^P \sum_{i=1}^{I_p} a_{wpi} n_{pi} = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2},$$

and

$$\mathbf{b}_w = (b_{wp})_{p=1,\dots,P} \text{ and } b_{wp} = \frac{n_p \bar{w}_p^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2},$$

with

$$\bar{w}_p = \frac{1}{n_p} \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} = \frac{1}{n_p} \sum_{i=1}^{I_p} n_{pi} \bar{w}_{pi},$$

and

$$\bar{n}_{psu}(\mathbf{b}_w) = \sum_{p=1}^P b_{wp} n_p = \frac{\sum_{p=1}^P \left(\sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}.$$

Note that $\bar{n}_{int}(\mathbf{A}_w) \leq \bar{n}_{psu}(\mathbf{b}_w)$ with equality if and only if $I_p = 1$. Also note that $\bar{n}_{int}(\mathbf{A}_w)$ is invariant of the allocation of the interviewers to the psu's while $\bar{n}_{psu}(\mathbf{b}_w)$ is not.

Remark 4.1: If $\rho_C = 0$ we get

$$ieff_{s,w} = 1 + \rho_{int}(\bar{n}_{int}(\mathbf{A}_w) - 1).$$

This formula is similar to $ieff_w$ given in Section 2. Thus, all the comments given in Remark 2.1 apply here. Note that $\bar{n}_{int}(\mathbf{A}_w)$, just like \bar{n}_w , cannot be generally written in the form $\bar{n}_{int}(\mathbf{A}_w) = \sum_{p=1}^P \sum_{i=1}^{I_p} a_{wpi} n_{pi}$ with $\sum_{p=1}^P \sum_{i=1}^{I_p} a_{wpi} = 1$; the same comment applies to $\bar{n}_{psu}(\mathbf{b}_w)$.

Remark 4.2: Define

$$\bar{n}_{int}(\mathbf{A}) = \sum_{p=1}^P \sum_{i=1}^{I_p} a_{pi} n_{pi}, \text{ where } \mathbf{A} = ((a_{pi})), \text{ with } a_{pi} = \frac{n_{pi}}{n},$$

and

$$\bar{n}_{psu}(\mathbf{b}) = \sum_{p=1}^P b_p n_p, \text{ where } \mathbf{b} = (b_1, \dots, b_P) \text{ with } b_p = \frac{n_p}{n}.$$

If $\rho_C \neq 0$ but we have an epsem design, then we drop the suffix w in $ieff_{s,w}$. Note that

$$\begin{aligned} ieff_s &= 1 + \rho_{int} \frac{\bar{n}_{int}(\mathbf{A}) - 1}{1 + \rho_C [\bar{n}_{psu}(\mathbf{b}) - 1]} \\ &= 1 + \frac{\rho_{int}}{\rho_C} \cdot \frac{\bar{n}_{int}(\mathbf{A}) - 1}{\bar{n}_{psu}(\mathbf{b}) - 1} \cdot \frac{\rho_C (\bar{n}_{psu}(\mathbf{b}) - 1)}{1 + \rho_C (\bar{n}_{psu}(\mathbf{b}) - 1)} \end{aligned}$$

so that

$$ieff_s < 1 + \frac{\rho_{int}}{\rho_C} \cdot \frac{\bar{n}_{int}(\mathbf{A}) - 1}{\bar{n}_{psu}(\mathbf{b}) - 1} < 1 + \frac{\rho_{int}}{\rho_C} \cdot \frac{\bar{n}_{int}(\mathbf{A})}{\bar{n}_{psu}(\mathbf{b})}.$$

It can be readily seen that the right side of the inequality increases with the ratios ρ_{int}/ρ_C and

$$\frac{\bar{n}_{int}(\mathbf{A}) - 1}{\bar{n}_{psu}(\mathbf{b}) - 1}.$$

We have

$$ieff_s - ieff = \rho_{int} \frac{\frac{\bar{n}_{int}(\mathbf{A}) - 1}{\bar{n}_{int} - 1} - [1 + \rho_C (\bar{n}_{psu}(\mathbf{b}) - 1)]}{1 + \rho_C (\bar{n}_{psu}(\mathbf{b}) - 1)} (\bar{n}_{int} - 1).$$

Thus, for $\rho_{int} > 0$,

$ieff_s < ieff$ if and only if

$$Deff_s := 1 + \rho_C (\bar{n}_{psu}(\mathbf{b}) - 1) > \frac{\bar{n}_{int}(\mathbf{A}) - 1}{\bar{n}_{int} - 1},$$

i.e., if and only if the design effect due to the spatial clustering is larger than the ratio of the weighted average of the interviewer workload -1 and the average interviewer workload -1 . If the interviewer workload is the same for all the interviewers, the right hand side of the inequality is 1 and so the inequality is always valid. It is interesting to note that $ieff \approx 4 \cdot ieff_s$ if $\rho_{int} = 0.1$, $\rho_C = 0.05$, $\bar{n}_{psu}(\mathbf{b}) = 140$, and $\bar{n}_{int} = 70$.

Remark 4.3: In the general case, we have

$$ieff_{s,w} - ieff = \rho_{int} \left(\frac{\bar{n}_{int}(\mathbf{A}_w) - 1}{1 + \rho_C (\bar{n}_{psu}(\mathbf{b}_w) - 1)} - (\bar{n}_{int} - 1) \right).$$

Thus, for $\rho_{int} > 0$,

$ieff_{s,w} < ieff$ if and only if

$$Deff_{s,w} := 1 + \rho_C (\bar{n}_{psu}(\mathbf{b}_w) - 1) > \frac{\bar{n}_{int}(\mathbf{A}_w) - 1}{\bar{n}_{int} - 1},$$

i.e., if and only if

$$\rho_C > \frac{\bar{n}_{int}(\mathbf{A}_w) - \bar{n}_{int}}{(\bar{n}_{int} - 1)(\bar{n}_{psu}(\mathbf{b}_w) - 1)} =: \rho_C^*, \text{ say.}$$

In Example 2 (see Table 3), $ieff$ is a conservative value for $ieff_{s,w}$ for a) to e) if $\rho_C > 0$. The same holds for f) to h) if $\rho_C > 0.004$.

Table 3
Average interviewer workloads for several parameter combinations (Example 2); $ieff / ieff_{s,w}$ for $\rho_{int} = 0.01$ and $\rho_c = 0.02$

	n_i	\bar{w}_i	σ_i^2	$IA = (1,3)$				$IA = (2,2)$				$IA = (3,1)$				
				\bar{n}_{int}	$\bar{n}_{int}(A_w)$	$\bar{n}_{psu}(b_w)$	ρ_c^*	$\frac{ieff}{ieff_{s,w}}$	$\bar{n}_{int}(A_w)$	$\bar{n}_{psu}(b_w)$	ρ_c^*	$\frac{ieff}{ieff_{s,w}}$	$\bar{n}_{int}(A_w)$	$\bar{n}_{psu}(b_w)$	ρ_c^*	$\frac{ieff}{ieff_{s,w}}$
a)	25	1.022	0.299	25	19.202	47.528	-0.005	1.133	19.202	38.389	-0.006	1.123	19.202	49.039	-0.005	1.135
	25	1.036	0.375													
	25	0.998	0.276													
	25	0.945	0.260													
b)	10	1	1	25	15	41	-0.010	1.151	15	29	-0.015	1.138	15	26	-0.017	1.134
	20	1	1													
	30	1	1													
	40	1	1													
c)	10	1	1	25	7.5	20.5	-0.037	1.185	7.5	14.5	-0.054	1.180	7.5	13	-0.061	1.178
	20	1	2													
	30	1	3													
	40	1	4													
d)	10	1	4	25	10	27.333	-0.024	1.171	10	19.333	-0.034	1.163	10	17.333	-0.038	1.161
	20	1	3													
	30	1	2													
	40	1	1													
e)	10	4	144	25	1.801	2.755	-0.551	1.230	1.801	3.603	-0.371	1.231	1.801	4.344	-0.289	1.231
	20	2	9													
	30	0.333	0.555													
	40	0.250	0.125													
f)	10	0.333	0.025	25	31.820	75.685	0.004	1.104	31.820	58.427	0.005	1.084	31.820	40.629	0.007	1.058
	20	0.666	0.075													
	30	1	0.125													
	40	1.333	0.175													
g)	10	1	0.010	25	29.126	79.612	0.002	1.118	29.126	56.311	0.003	1.094	29.126	50.485	0.003	1.086
	20	1	0.020													
	30	1	0.030													
	40	1	0.040													
h)	10	1	0.004	25	29.940	81.836	0.003	1.117	29.940	57.884	0.004	1.092	29.940	51.896	0.004	1.084
	20	1	0.003													
	30	1	0.002													
	40	1	0.001													

If a household and a person within the household are selected at random, then the weights are often independent of the psu and the interviewer and depend only on the household sizes. In such a situation, the household sizes form the weighting classes. For weighting classes, we define

m_{pij} : number of sampling units in psu p assigned to interviewer i belonging to weighting class j ,

$m_{pj} = \sum_{i=1}^{I_p} m_{pij}$: number of sampling units in psu p belonging to weighting class j ,

$m_j = \sum_{p=1}^P \sum_{i=1}^{I_p} m_{pij}$: number of sampling units belonging to weighting class j .

Thus,

$n_{pi} = \sum_{j=1}^J m_{pij}$: number of sampling units in psu p assigned to interviewer i ,

$n_p = \sum_{i=1}^{I_p} \sum_{j=1}^J m_{pij}$: number of sampling units in psu p ,

$n = \sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{j=1}^J m_{pij}$: sample size.

Furthermore,

$$\bar{n}_{int}(A_w) = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{j=1}^J w_j m_{pij} \right)^2}{\sum_{j=1}^J w_j^2 m_j}$$

and

$$\bar{n}_{psu}(b_w) = \frac{\sum_{p=1}^P \left(\sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} = \frac{\sum_{p=1}^P \left(\sum_{j=1}^J w_j m_{pj} \right)^2}{\sum_{j=1}^J w_j^2 m_j},$$

are ratios of quadratic forms in $\mathbf{w} = (w_1, \dots, w_J)$.

5. Overall effects

The overall effects take into account unequal weighting, spatial clustering, and the interview effects and can be viewed as a generalization to the traditional design effects. Multiplying the SRS variance for the unweighted sample mean by the overall effects will provide the total variance estimator.

$$eff = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_1'}(\bar{y})} = eff_w \times eff_s \times eff_{int},$$

where

$$eff_w = \frac{\text{Var}_{M_1'}(\bar{y}_w)}{\text{Var}_{M_1'}(\bar{y})},$$

$$eff_s = \frac{\text{Var}_{M_3}(\bar{y}_w)}{\text{Var}_{M_1'}(\bar{y}_w)},$$

$$eff_{int} = ieff_{s,w} = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_3}(\bar{y}_w)}.$$

In the above, $\text{Var}_{M_1'}$ is with respect to the following model:

$$M_1': \text{Cov}(y_{pik}, y_{p'i'k'}) = \begin{cases} \sigma^2 & \text{if } p = p', i = i', k = k', \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that

$$eff = \frac{n \sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2} \times \left[1 + \rho_C \left(\frac{\sum_{p=1}^P \left(\sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} - 1 \right) + \rho_{int} \left(\frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} - 1 \right) \right].$$

The relative contributions of weighting, spatial clustering, and interviewer effects to the overall effects are given by

$$Reff_w = \frac{n \sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2},$$

$$Reff_s = \frac{1 + \rho_C(\bar{n}_{psu}(\mathbf{b}_w) - 1)}{eff},$$

$$Reff_I = \frac{1 + \rho_{int} \frac{\bar{n}_{psu}(A_w) - 1}{1 + \rho_C(\bar{n}_{psu}(\mathbf{b}_w) - 1)}}{eff}.$$

In Figure 2, we present three dimensional graphs of the relative contributions of weighting, spatial clustering, and interviewer effects to the overall effects for different combinations of intra-cluster and intra-interviewer correlations for different patterns of weights given in cases a), f) and h) of Table 3 with $IA = (1, 3)$, where $IA = (a, b)$ indicates that the first a of the four interviewers are in psu 1 and the last b interviewers are in psu 2.

Remark 5.1: From Result 6, we get

$$eff \geq 1 + \rho_C \left(\frac{n}{P} - 1 \right) + \rho_{int} \left(\frac{n}{I} - 1 \right).$$

The right side is the overall effect if the same number of interviewers with equal workload is assigned to each psu. It is interesting to note the similarity between the right hand side of the above inequality and the design effects formula given in (3.1) of Hansen, Hurwitz and Madow (1953, Vol. I, page 370). To claim the similarity, we need to treat the secondary sampling units as the units belonging to an interviewer. In this connection, we also note the formula (3.7) given in Hansen *et al.* (1953, Vol. II, page 292) for the case $I = P$.

Remark 5.2: When we have the same weighting classes across psu \times interviewer, we have

$$eff = \frac{n \sum_{j=1}^J w_j^2 m_j}{\left(\sum_{j=1}^J w_j m_j \right)^2} \times \left[1 + \rho_C \left(\frac{\sum_{p=1}^P \left(\sum_{j=1}^J w_j m_{pj} \right)^2}{\sum_{j=1}^J w_j^2 m_j} - 1 \right) + \rho_{int} \left(\frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{j=1}^J w_j m_{pij} \right)^2}{\sum_{j=1}^J w_j^2 m_j} - 1 \right) \right].$$

Remark 5.3: Consider the special case

$$m_{pij} = \frac{n_{pi} m_j}{n}$$

in which we allow variation in weights within psu \times interviewer classes, but we constrain the weights to have the same relative frequency distribution in each class, *i.e.*, the means and the variances of the weights within the classes do not depend on the class (Lynn and Gabler 2004). It is easy to see that in this case

$$eff = \frac{n \sum_{j=1}^J w_j^2 m_j}{\left(\sum_{j=1}^J w_j m_j \right)^2} \times \left[1 + \rho_C \left(\frac{\left(\sum_{j=1}^J w_j m_j \right)^2 \sum_{p=1}^P n_p^2}{\sum_{j=1}^J w_j^2 m_j n^2} - 1 \right) + \rho_{int} \left(\frac{\left(\sum_{j=1}^J w_j m_j \right)^2 \sum_{p=1}^P \sum_{i=1}^{I_p} \frac{n_{pi}^2}{n^2} - 1 \right) \right].$$

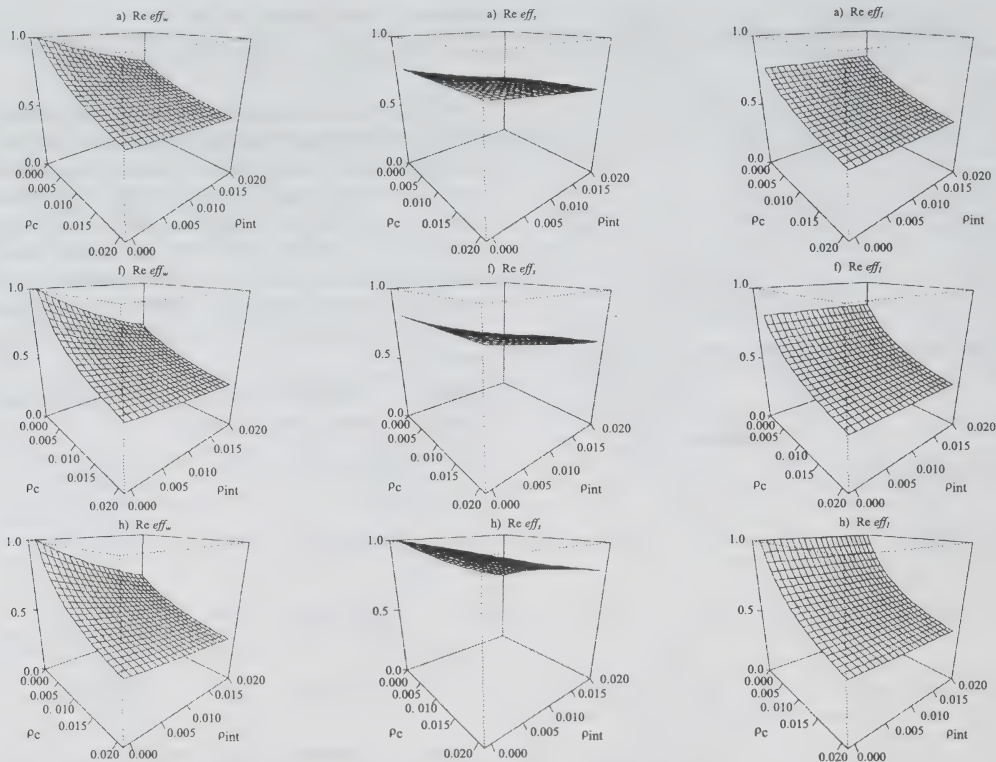


Figure 2 Relative contributions of weighting, design and interviewer effects to the overall effects for cases a), f) and h) Example 2 for the case $IA = (1, 3)$

Using the same argument given in the proof of Result 6, we get

$$\begin{aligned} eff &\geq 1 + \rho_c \left(\frac{\sum_{p=1}^P n_p^2}{n^2} - 1 \right) + \rho_{int} \left(\frac{\sum_{p=1}^P \sum_{i=1}^{I_p} n_{pi}^2}{n} - 1 \right) \\ &= 1 + \rho_c (\bar{n}_{psu}(\mathbf{b}) - 1) + \rho_{int} (\bar{n}_{int}(\mathbf{A}) - 1). \end{aligned}$$

This means that the overall effect is larger than the overall effect for an epsem design (see Remark 5.4).

Remark 5.4: For an epsem design, we have

$$eff = 1 + \rho_c (\bar{n}_{psu}(\mathbf{b}) - 1) + \rho_{int} (\bar{n}_{int}(\mathbf{A}) - 1),$$

where

$$\bar{n}_{psu}(\mathbf{b}) = \frac{\sum_{p=1}^P n_p^2}{n} \text{ and } \bar{n}_{int}(\mathbf{A}) = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} n_{pi}^2}{n}.$$

Note that Davis and Scott (1995) obtained this formula for the special case of the following linear mixed model:

$$y_{pik} = \mu + \alpha_i + \beta_p + \varepsilon_{pik},$$

where μ is the overall effect, α_i, β_p are random effects due to the interviewer i , psu p and ε_{pik} is the pure error. They assumed that the random effects are independent with $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_p \sim N(0, \sigma_\beta^2)$ and $\varepsilon_{pik} \sim N(0, \sigma_\varepsilon^2)$.

For the above linear mixed model, it is easy to check that

$$\rho_{int} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2} \text{ and } \rho_c = \frac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2}.$$

However, it is instructive to note that the definition eff does not require ρ_{int} and ρ_c to be strictly positive and the definition goes beyond the linear mixed model. For example, the definition applies to the following example:

Example 3: A simple model for binary data.

Assuming $0 < \min(\alpha, \beta) < \theta < 1$, we define the following model:

For all n_{pi} different respondents of interviewer i in psu p .

$$P(Y_{pik} = x_1, Y_{pi'k'} = x_2)$$

$x_1 \backslash x_2$	1	0	Total
1	α	$\theta - \alpha$	θ
0	$\theta - \alpha$	$1 - 2\theta + \alpha$	$1 - \theta$
Total	θ	$1 - \theta$	1

For all n_{pi} respondents of interviewer i and psu p and all $n_{pi'}$ respondents of interviewer i' and psu p .

$$P(Y_{pik} = x_1, Y_{pi'k'} = x_2)$$

$x_1 \backslash x_2$	1	0	Total
1	β	$\theta - \alpha$	θ
0	$\theta - \alpha$	$1 - 2\theta + \beta$	$1 - \theta$
Total	θ	$1 - \theta$	1

For all n_p respondents of psu p and all $n_{p'}$ respondents of psu p' .

$$P(Y_{pik} = x_1, Y_{p'i'k'} = x_2)$$

$x_1 \backslash x_2$	1	0	Total
1	θ^2	$\theta(1 - \theta)$	θ
0	$\theta(1 - \theta)$	$(1 - \theta)^2$	$1 - \theta$
Total	θ	$1 - \theta$	1

Therefore, we have

$$E(Y_{pik}) = \theta \text{ for all } p, i, k,$$

$$\text{Var}(Y_{pik}) = \theta(1 - \theta) \text{ for all } p, i, k,$$

$$\begin{aligned} \rho &= \frac{\text{Cov}(Y_{pik}, Y_{pi'k'})}{\sqrt{\text{Var}(Y_{pik})\text{Var}(Y_{pi'k'})}} \\ &= \frac{\alpha - \theta^2}{\theta(1 - \theta)} \text{ for all } p, i \text{ and } k \neq k', \end{aligned}$$

$$\begin{aligned} \rho_C &= \frac{\text{Cov}(Y_{pik}, Y_{p'i'k'})}{\sqrt{\text{Var}(Y_{pik})\text{Var}(Y_{p'i'k'})}} \\ &= \frac{\beta - \theta^2}{\theta(1 - \theta)} \text{ for all } p \text{ and } i \neq i', \end{aligned}$$

which is a special case of Model M_4 with $\sigma^2 = \text{Var}(Y_{pik}) = \theta(1 - \theta)$. Note that both ρ_C and ρ may be negative and $\rho_{\text{int}} = \rho - \rho_C$ is positive if and only if $\alpha > \beta$.

Remark 5.5: For an epsem design with common psu size $b = n/P$, we have

$$\text{eff} = 1 + \rho_C(b - 1) + \rho_{\text{int}}(\bar{n}_{\text{int}}(A) - 1).$$

Remark 5.6: In discussing Verma *et al.* (1980), Holt considered the case when there is no interviewer variability and psu is the weighting class, i.e., the case when $\rho_{\text{int}} = 0$ and $w_{pik} = w_p$ for all p, i, k . In this case *eff* reduces to

$$\text{eff} = \frac{n \sum_{p=1}^P n_p w_p^2}{\left(\sum_{p=1}^P n_p w_p \right)^2} \times \left[1 + \rho_C \left(\frac{\sum_{p=1}^P n_p^2 w_p^2}{\sum_{p=1}^P n_p w_p^2} - 1 \right) \right].$$

Note that the above formula can be obtained from equation (A4) of Holt in discussing Verma *et al.* (1980), after correcting an obvious typo (i.e., deleting n in the denominator), choosing his choice of survey weight and some algebra. Design effect formulae in the absence of the interviewer effects were considered by many authors. See Kish (1965), Verma *et al.* (1980), Skinner (1986), Valliant (1987), Skinner *et al.* (1989), Gabler, Häder and Lahiri (1999), Lynn and Gabler (2004), Kalton, Brick and Lé (2005) and others.

6. Concluding remarks

We have noticed that the standard interviewer effects formula could have either an overestimation or underestimation problem depending on the situation. For example, it could severely underestimate the interviewer effects in an epsem sampling design with different interviewer workloads. Interestingly, spatial correlation can turn this underestimation to an overestimation. In the former case, the survey designer who uses the standard interviewer effect formula may pay little attention to control the interviewer effect. In the latter case, a high value of the interviewer effect may unnecessarily raise concerns about the quality of data connected with the interviewer. This may trigger allocation of a higher portion of budget than is necessary to reduce the interviewer effect, which may be already much lower than the value obtained by an application of the standard formula. The paper is an attempt to define and interpret interviewer effects that are appropriate in different complex survey situations.

We have considered the case when an interviewer is assigned only in one psu. The case when an interviewer works in different psu's is also important and will be considered in a later paper. The weights used in the proposed formulae only account for sampling weights as they are planned at the design stage, but do not necessarily reflect the actual weights attached to each case once the data are collected. In other words, our interviewer effect formulae do not incorporate the effects due to nonresponse and post-stratification adjustments. The formulae presented in the paper are mainly useful in the planning and design stage when we have some ideas about the intra-interviewer and spatial correlations.

Reliable estimation of ρ_{int} and ρ_c is important. Although there are some papers that deal with the estimation of ρ_{int} and ρ_c , there is certainly a need to advance research in this important area. In comparing the two sources of homogeneity, Hansen *et al.* (1961) found that the interviewer variability was often larger than the sampling variability. In many surveys, such an evaluation, which requires estimation of the intra-interviewer and intra-cluster correlations, is either difficult or even impossible because the interviewer effects are often confounded with the spatial clustering effects. The use of an interpenetrating design, first proposed by Mahalanobis (1946), where respondents are randomly assigned to the interviewers, is a way to get around the problem. In practice, the implementation of such a design in a large scale sample survey is difficult, but some approximated interpenetrated designs can be applied (Hansen *et al.* 1961, Bailer, Bailey and Stevens 1977, Bailey, Moore and Bailer 1978, Collins and Butcher 1982, O'Muircheartaigh and Campanelli 1998). Multi-level models have been used as a partial remedy to the problem (Hox and De Leeuw 1994, Davis and Scott 1995, O'Muircheartaigh and Campanelli 1998, Scott and Davis 2001). We have not considered the problem of the estimation of the intra-interviewer and intra-cluster correlations. This is an important problem and will be considered in a later paper.

In practice, interviewer or design effects are computed for many items using the same formula and a summary measure such as the median interviewer or design effect is taken for the planning and design of the survey. So far as the issues related to handling multiple items are concerned, one may continue to follow one's own protocol; the only change we may suggest is to use our new definitions for interviewer effects or overall effects whenever applicable. The use of our formula may suggest overall effects, which may be much lower than the standard formula. This, in turn, may suggest lower sample size and hence may save survey costs.

Appendix

$$\text{Result 1. } ieff_w = \frac{\text{Var}_{M_2}(\bar{y}_w)}{\text{Var}_{M_1}(\bar{y}_w)} = 1 + \rho_{\text{int}} \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right).$$

Proof: The result follows by noting

$$\text{Var}_{M_1}(\bar{y}_w) = \text{Var}_{M_1} \left[\frac{\sum_i \sum_k w_{ik} y_{ik}}{\sum_i \sum_k w_{ik}} \right] = \frac{\sigma^2 \sum_i \sum_k w_{ik}^2}{\left(\sum_i \sum_k w_{ik} \right)^2},$$

and

$$\text{Var}_{M_2}(\bar{y}_w) = \frac{\sigma^2 \left[\sum_i \sum_k w_{ik}^2 + \rho_{\text{int}} \sum_i \sum_{k \neq k'} w_{ik} w_{ik'} \right]}{\left(\sum_i \sum_k w_{ik} \right)^2},$$

and some algebra.

Corollary: Assume $\rho_{\text{int}} > 0$ and $w_{ik} = 1/n$. Using Result 1 and the Cauchy-Schwarz inequality, we get

$$ieff(\mathbf{a}_1) = 1 + \rho_{\text{int}} \left(\frac{\sum_i n_i^2}{n} - 1 \right) \geq 1 + \rho_{\text{int}} \left(\frac{n}{I} - 1 \right) = ieff.$$

Result 2. $ieff_w \leq ieff(\mathbf{a}_2)$, where

$$\mathbf{a}_2 = (a_{21}, \dots, a_{2I}) \text{ with } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2}.$$

Proof: Using the Cauchy-Schwarz inequality, we have

$$\sum_i \left(\sum_k w_{ik} \right)^2 \leq \sum_i n_i \sum_k w_{ik}^2$$

with equality if and only if $w_{ik} = \bar{w}_i$ for all i and k , where

$$\bar{w}_i = \frac{\sum_{k=1}^{n_i} w_{ik}}{n_i}$$

is the average survey weight for the i^{th} interviewer. Thus, we have $ieff_w \leq 1 + [\bar{n}_{\text{int}}(\mathbf{a}_2) - 1] \rho_{\text{int}} = ieff(\mathbf{a}_2)$.

The equality holds if and only if $w_{ik} = \bar{w}_i$ for all i and k in which case $ieff_w = ieff(\mathbf{a}_2^*)$, where

$$\mathbf{a}_2^* = (a_{21}^*, \dots, a_{2I}^*), \text{ with } a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}.$$

If all weights are non-negative, then

$$\sigma_i^2 = \frac{1}{n_i} \sum_k (w_{ik} - \bar{w}_i)^2 \leq (n_i - 1) \bar{w}_i^2,$$

since σ_i^2 is Schur-convex. Defining

$$x_i = \frac{1 + \frac{\sigma_i^2}{\bar{w}_i^2}}{n_i} \text{ implies } \frac{1}{n_i} \leq x_i \leq 1$$

and

$$\begin{aligned} \bar{n}_{\text{int}}(\mathbf{a}_2) &= \frac{\sum_i n_i \sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2} = \frac{\sum_i n_i^2 \bar{w}_i^2 + \sum_i n_i^2 \sigma_i^2}{\sum_i n_i \bar{w}_i^2 + \sum_i n_i \sigma_i^2} \\ &= \frac{\sum_i n_i^3 \bar{w}_i^2 - \sum_i n_i^2 ((n_i - 1) \bar{w}_i^2 - \sigma_i^2)}{\sum_i n_i^2 \bar{w}_i^2 - \sum_i n_i ((n_i - 1) \bar{w}_i^2 - \sigma_i^2)} \\ &= \frac{\sum_i n_i^3 \bar{w}_i^2 x_i}{\sum_i n_i^2 \bar{w}_i^2 x_i} \leq \frac{\sum_i n_i^3 \bar{w}_i^2}{\sum_i n_i^2 \bar{w}_i^2} = \sum_i n_i \frac{n_i^2 \bar{w}_i^2}{\sum_i n_i^2 \bar{w}_i^2} \end{aligned}$$

with equality if and only if $\sigma_i^2 = (n_i - 1) \bar{w}_i^2$ for all i or if all n_i are equal.

The inequality follows from the logarithmic concavity of $\bar{n}_{\text{int}}(\mathbf{a}_2)$ as function of (x_1, \dots, x_I) .

Result 3. For $\mathbf{a}_2^* = (a_{21}^*, \dots, a_{2I}^*)$ with $a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}$

and

$$\mathbf{a}_2 = (a_{21}, \dots, a_{2I}) \text{ with } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2},$$

we have

$$\begin{aligned} \text{ieff}(\mathbf{a}_2^*) &\leq \text{ieff}(\mathbf{a}_2) \text{ if and only if} \\ &\geq \sum_i n_i \sigma_i^2 \sum_i n_i^2 \bar{w}_i^2 \leq \sum_i n_i^2 \sigma_i^2 \sum_i n_i \bar{w}_i^2. \end{aligned}$$

Proof. We have

$$\text{ieff}(\mathbf{a}_2^*) - \text{ieff}(\mathbf{a}_2) = \frac{\sum_i n_i \sigma_i^2 \sum_i n_i^2 \bar{w}_i^2 - \sum_i n_i^2 \sigma_i^2 \sum_i n_i \bar{w}_i^2}{\left(\sum_i n_i \sigma_i^2 + \sum_i n_i \bar{w}_i^2 \right) \sum_i n_i \bar{w}_i^2}.$$

For $n_i = n/I$ for all i , we get

$$\text{ieff}(\mathbf{a}_2^*) = \text{ieff}(\mathbf{a}_2).$$

For $w_{ik} = \bar{w}_i$ for all i , i.e., $\sigma_i^2 = 0$, we get

$$\text{ieff}(\mathbf{a}_2^*) = \text{ieff}(\mathbf{a}_2).$$

For $\bar{w}_i = \bar{w}$ for all i and $\sigma_i^2 = \sigma^2 > 0$ for all i , we get

$$\text{ieff}(\mathbf{a}_2^*) = \text{ieff}(\mathbf{a}_2).$$

For $\bar{w}_i = \bar{w}$ for all i , we get

$$\text{ieff}(\mathbf{a}_2^*) \leq \text{ieff}(\mathbf{a}_2) \text{ iff } \sum_i n_i \sigma_i^2 \sum_i n_i^2 \leq \sum_i n_i^2 \sigma_i^2.$$

For $\sigma_i^2 = \sigma^2 > 0$ for all i , we get

$$\text{ieff}(\mathbf{a}_2^*) \leq \text{ieff}(\mathbf{a}_2) \text{ iff } n \sum_i n_i^2 \bar{w}_i^2 \leq \sum_i n_i \bar{w}_i^2 \sum_i n_i^2.$$

Result 4. We have

$$\begin{aligned} \text{ieff}_w - \text{ieff} &= \frac{\bar{n}_{\text{int}}}{SST + n\bar{w}^2} \left[\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SSW \right] \rho_{\text{int}} \\ &= \frac{\bar{n}_{\text{int}}}{(1 + CV_w^{-2})SST} \left[\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SSW \right] \rho_{\text{int}} \\ &= \frac{\bar{n}_{\text{int}} \tau_w}{1 + CV_w^{-2}} \left(\frac{\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2}{SSW} - 1 \right) \rho_{\text{int}}. \end{aligned}$$

Proof.

$$\text{ieff}_w - \text{ieff}$$

$$= 1 + \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right) \rho_{\text{int}} - 1 - (\bar{n}_{\text{int}} - 1) \rho_{\text{int}}$$

$$= \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - \bar{n}_{\text{int}} \right) \rho_{\text{int}}$$

$$= \left(\frac{\sum_i n_i^2 \bar{w}_i^2}{SST + n\bar{w}^2} - \bar{n}_{\text{int}} \right) \rho_{\text{int}}$$

$$= \frac{\bar{n}_{\text{int}}}{SST + n\bar{w}^2} \left(\frac{\sum_i n_i^2 \bar{w}_i^2}{\bar{n}_{\text{int}}} - (SST + n\bar{w}^2) \right) \rho_{\text{int}}$$

$$= \frac{\bar{n}_{\text{int}}}{SST + n\bar{w}^2} \left(\sum_{i=1}^I \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 + \sum_{i=1}^I n_i \bar{w}_i^2 - (SST + n\bar{w}^2) \right) \rho_{\text{int}}.$$

Now the result follows using algebra.

Result 5.

$$ieff_{s,w} = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_3}(\bar{y}_w)}$$

$$= 1 + \rho_{\text{int}} \frac{\left(\frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} - 1 \right)}{\left(\frac{\sum_{p=1}^P \left(\sum_{i=1}^{I_p} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2 \right)}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} - 1 \right)}.$$

Proof. The result follows by noting that

$$\frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_3}(\bar{y}_w)} =$$

$$\frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2 + \rho_C \sum_{p=1}^P \sum_{i \neq i'} \sum_{k=1}^{n_{pi}} \sum_{k'=1}^{n_{pi'}} w_{pik} w_{pi'k'} + \rho \sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k \neq k'} \sum_{k'=1}^{n_{pi}} w_{pik} w_{pi'k'}}{\sum_{p=1}^P \left(\sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2 + \rho_C \sum_{i,i'} \sum_{k \neq k'} w_{pik} w_{pi'k'} \right)}$$

and some algebra.

Result 6. For $0 < \rho_C < 1$ and $0 < \rho_{\text{int}} < 1$,

$$eff \geq 1 + \rho_C \left(\frac{n}{P} - 1 \right) + \rho_{\text{int}} \left(\frac{n}{I} - 1 \right),$$

with equality if and only if the weights are all equal and each interviewer has the same workload.

If we have in each psu only one interviewer, then

$$eff \geq 1 + (\rho_C + \rho_{\text{int}}) \left(\frac{n}{P} - 1 \right).$$

Proof. Using some algebra and the general inequality,

$$\sum_j p_j x_j^2 \geq \left(\sum_j p_j x_j \right)^2$$

with

$$p_j \geq 0 \text{ and } \sum_{j=1}^J p_j = 1,$$

we have

$$\begin{aligned} eff &= \frac{n \sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2} (1 - \rho_C - \rho_{\text{int}}) \\ &\quad + n \rho_C \frac{\sum_{p=1}^P \left(\sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\left(\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2} + n \rho_{\text{int}} \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\left(\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2} \\ &\geq 1 - \rho_C - \rho_{\text{int}} + \rho_C \frac{n}{P} + \rho_{\text{int}} n \frac{\sum_{p=1}^P \frac{I_p}{I} \left(\frac{1}{I_p} \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\left(\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} \right)^2} \end{aligned}$$

$$\geq 1 - \rho_C - \rho_{\text{int}} + \rho_C \frac{n}{P} + \rho_{\text{int}} \frac{n}{I}$$

$$= 1 + \rho_C \left(\frac{n}{P} - 1 \right) + \rho_{\text{int}} \left(\frac{n}{I} - 1 \right).$$

Acknowledgements

We thank the editor and referees for constructive comments and suggestions which have improved the original version of this paper.

References

- Bailar, B.A., Bailey, L. and Stevens, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.
- Bailey, L., Moore, T.F. and Bailar, B.A. (1978). An interviewer variance study for eight impact cities of the National Crime Survey Cities Sample. *Journal of the American Statistical Association*, 73, 16-23.
- Biemer, P.P., and Stokes, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 369, 158-166.
- Biemer, P., and Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 603-632.
- Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, 5, 93-105.
- Chambers, R.L., and Skinner, C.J. (Eds.) (2003). *Analysis of Survey Data*. Wiley, Chichester.

- Collins, M., and Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.
- Davis, P.D., and Scott, A.J. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, 21, 99-106.
- Feather, J. (1973). A study of interviewer variance. WHO International Collaborative Study of Medical Care Utilization, Saskatchewan Study Area Reports, Series II, Monograph No. 3.
- Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.
- Gray, P.G. (1956). Examples of interviewer variability taken from two sample surveys. *Applied Statistics*, V, 73-85.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- Groves, R.M., and Magilav, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.
- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol I, II. New York: John Wiley & Sons, Inc.
- Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961). Measurement errors in census and surveys. *Bulletin of the ISI* 38, 2, 351-374.
- Hanson, R.H., and Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.
- Hájek, J. (1971). Comments, In Foundations of Statistical Inference, (Eds. V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart, and Winston.
- Heeb, J.-L., and Gmel, G. (2001). Interviewers and respondents effects on self-reported alcohol consumption in Swiss Health Survey. *Journal of Studies on Alcohol*, 62, 434-442.
- Hox, J.J., and De Leeuw, E.D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modeling to meta-analysis. *Quality & Quantity*, 329-344.
- Kalton, G., Brick, J.M. and Lê, T.h. (2005). Estimating components of design effects for use in sample design. In: *Household Sample Surveys in Developing and Transition Countries*, Chapter VI. Available from http://unstats.un.org/unsd/hhsurveys/pdf/Chapter_6.pdf.
- Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Lynn, P., and Gabler, S. (2004). Approximations to b^* in the estimation of design effects due to clustering. *Working Papers of the Institute for Social and Economic Research*, paper 2004-07. Colchester: University of Essex. Available from <http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-07.pdf>.
- Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, Serie A, 109, 325-378, reprinted in *Sankhyā* (1958), 1-68.
- O'Muircheartaigh, C., and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society*, Series A, 161, 63-77.
- Philippens, M., and Loosveldt, G. (2004). Interviewer-related variance in the European Social Survey. Paper presented at the sixth international conference on social science methodology, August 17-20 in Amsterdam.
- Rice, S.A. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology*, 35, 420-423.
- Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multi-way contingency tables with proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- Schnell, R., and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.
- Scott, A.J., and Davis, P.D. (2001). Estimating interviewer effects for binary responses. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency*.
- Skinner, C.J. (1986). Design effect of two-stage sampling, *Journal of the Royal Statistical Society*, Serie B, 48, 89-99.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.
- Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47, 84-95.
- Valliant, R.M. (1987). Generalized variance functions in stratified two-stage sampling, 82, 499-508.
- Verma, V., Scott, C. and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Serie A, 143, 431-473.

Indicators for the representativeness of survey response

Barry Schouten, Fannie Cobben and Jelke Bethlehem¹

Abstract

Many survey organisations focus on the response rate as being the quality indicator for the impact of non-response bias. As a consequence, they implement a variety of measures to reduce non-response or to maintain response at some acceptable level. However, response rates alone are not good indicators of non-response bias. In general, higher response rates do not imply smaller non-response bias. The literature gives many examples of this (*e.g.*, Groves and Peytcheva 2006, Keeter, Miller, Kohut, Groves and Presser 2000, Schouten 2004).

We introduce a number of concepts and an indicator to assess the similarity between the response and the sample of a survey. Such quality indicators, which we call R-indicators, may serve as counterparts to survey response rates and are primarily directed at evaluating the non-response bias. These indicators may facilitate analysis of survey response over time, between various fieldwork strategies or data collection modes. We apply the R-indicators to two practical examples.

Key Words: Quality; Non-response; Non-response reduction; Non-response adjustment.

1. Introduction

It is a well-developed finding in the survey methodological literature that response rates by themselves are poor indicators of non-response bias, see *e.g.*, Curtin, Presser and Singer (2000), Groves, Presser and Dipko (2004), Groves (2006), Groves and Peytcheva (2006), Keeter *et al.* (2000), Merkle and Edelman (2002), Heerwegh, Abts and Loosveldt (2007) and Schouten (2004). However, the field has yet to propose alternative indicators of non-response that may be less ambiguous as indicators of survey quality.

We propose an indicator, which we call an R-indicator ('R' for representativeness), for the similarity between the response to a survey and the sample or the population under investigation. This similarity can be referred to as "representative response". In the literature, there are many different interpretations of the 'representativeness' concept. See Kruskal and Mosteller (1979a, b and c) for a thorough investigation of the statistical and non-statistical literature. Rubin (1976) introduced the concept of ignorable non-response; the minimal conditions that allow for unbiased estimation of a statistic. Some authors explicitly define representativeness. Hájek (1981) links "representative" to the estimation of population parameters; the pair formed by an estimator and a missing-data mechanism are representative when, with probability one, the estimator is equal to the population parameter. Following Hájek's definition, calibration estimators (*e.g.*, Särndal, Swensson and Wretman 2003) are representative for the auxiliary variables that are calibrated. Bertino (2006) defines a so-called univariate representativeness index for continuous random variables. This index is a distribution-free measure based on the Cramér – Von Mises statistic. Kohler (2007) defines what he calls an internal criterion for representativeness. His

univariate criterion resembles the Z-statistic for population means.

We separate the concept of representativeness from the estimation of a specific population parameter but relate this concept to the impact on the overall composition of response. By separating indicators from a specific parameter, they can be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. Also, the measure gives a multivariate perspective of the dissimilarity between sample and response.

The R-indicator that we propose employs estimated response probabilities. The estimation of response probabilities implies that the R-indicator itself is a random variable, and, consequently, has a precision and possibly a bias. The sample size of a survey, therefore, plays an important role in the assessment of the R-indicator as we will show. However, this dependence exists for any measure; small surveys simply do not allow for strong conclusions about the missing-data mechanism.

We show that the proposed R-indicator relates to Cramér's V measure for the association between response and auxiliary variables. In fact, we view the R-indicator as a lack-of-association measure. The weaker the association the better, as this implies there is no evidence that non-response has affected the composition of the observed data.

In order to be able to use R-indicators as tools for monitoring and comparing survey quality in the future, they need to have the features of a measure. That is, we want an R-indicator to be interpretable, measurable, able to be normalized and also to satisfy the mathematical properties of a measure. Especially since the interpretation and normalization are not straightforward features.

We apply the R-indicator to two studies that were conducted at Statistics Netherlands in 2005 and 2006. The objectives of those studies were the comparison of different data collection strategies. The studies involved different data collection modes and different non-response follow-up strategies. For each of the studies, a detailed analysis was done and documented. These studies are, therefore, suited to an empirical validation of the R-indicator. We compare the values of the R-indicator to the conclusions in the analyses. We refer to Schouten and Cobben (2007) and Cobben and Schouten (2007) for more illustrations and empirical investigations.

In section 2, we start with a discussion of the concept of representative response. Next, in section 3, we define the mathematical notation for our R-indicator. Section 4 is devoted to the features of the R-indicator. Section 5 describes the application of the R-indicator to the field studies. Finally, section 6 contains a discussion.

2. The concept of representative response

We, first, discuss what it means when a survey respondent pool is representative of the sample. Next, we make the concept of representativeness mathematically rigorous by giving it a definition.

2.1 What does representative mean?

Literature warns us not to single-mindedly focus on response rates as an indicator of survey quality. This can easily be illustrated by an example from the 1998 Dutch survey POLS (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English).

Table 1 contains the one- and two-month POLS survey estimates for the proportion of the Dutch population that receives a form of social allowance and the proportion that has at least one parent that was born outside the Netherlands. Both variables are taken from registry data and are artificially treated as survey items by deleting their values for non-respondents. The sample proportions are also given in Table 1. After one month, the response rate was 47.2%, while after the full two-month interview period, the rate was 59.7%. In the 1998 POLS, the first month was CAPI (Computer Assisted Personal Interview). Non-respondents after the first month were allocated to CATI (Computer Assisted Telephone Interview) when they had a listed, land-line phone. Otherwise, they were allocated once more to CAPI. Hence, the second interview month gave another 12.5% of response. However, from table 1 we can see that

after the second month, the survey estimates have a larger bias than after the first month.

Table 1
Response means in POLS for the first month of interviews and the full two-month interview period

Variable	After 1 month	After 2 months	Sample
Receiving social allowance	10.5%	10.4%	12.1%
Non-native	12.9%	12.5%	15.0%
Response rate	47.2%	59.7%	100%

From the example, it seems clear that the increased effort led to a less representative response with respect to both auxiliary variables. But what do we mean by representative in general?

It turns out that the term “representative” is often used with hesitation in the statistical literature. Kruskal and Mosteller (1979a, b and c) make an extensive inventory of the use of the word “representative” in the literature and identify nine interpretations. A number of interpretations they have found are omnipresent in the statistical literature. The statistical interpretations that Kruskal and Mosteller named ‘absence of selective forces’, ‘miniature of the population’, and ‘typical or ideal cases’ relate to probability sampling, quota sampling and purposive sampling. In the next section, we will propose a definition that corresponds to the ‘absence of selective forces’ interpretation. First, we will explain why we make this choice.

The concept of representative response is also closely related to the missing-data mechanisms Missing-Completely-at-Random (MCAR), Missing-at-Random (MAR) and Not-Missing-at-Random (NMAR) that are often referred to in the literature, see Little and Rubin (2002). A missing-data mechanism is MCAR when the probability of response does not depend on the survey topic of interest. The mechanism is MAR if the response probability depends on observed data only, which is, hence, a weaker assumption than MCAR. If the probability depends on missing data also, then the mechanism is said to be NMAR. These mechanisms, in fact, find their origin in model-based statistical theory. Somewhat loosely interpreted with respect to a survey topic, MCAR means that respondents are on average the same as non-respondents, MAR means that within known subpopulations, respondents are on average the same as non-respondents, and NMAR implies that even within subpopulations, respondents are different. The addition of the survey topic is essential. Within one questionnaire, some survey items can be MCAR, while other items are MAR or NMAR. Furthermore, the MAR assumption for one survey item holds for a particular stratification of the population. A different item may need a different stratification.

Given that we wish to monitor and compare the response to different surveys in topic or time, it is not appealing to define a representative response as dependent on the survey topic itself nor as dependent on the estimator used. We focus instead on the quality of data collection and not on the estimation. This setting leads us to compare the response composition to that of the sample. Clearly, the survey topics influence the probability that households participate in the survey, but the influence cannot be measured or tested and, hence, from our perspective, this influence cannot be the input for assessing response quality. We propose to judge the composition of response by pre-defined sets of variables that are observed outside of the survey and can be employed for each survey under investigation. We want the respondent selection to be as close as possible to a 'simple random sample of the survey sample', *i.e.*, with as little relation as possible between response and characteristics that distinguish units from each other. The latter can be interpreted as having selective forces which are absent in the selection of respondents, or as MCAR with respect to all possible survey variables.

2.2 Definition of a representative response subset

Let $i = 1, 2, 3, \dots, N$ be the unit labels for the population. By s_i we denote the 0-1-sample indicator, *i.e.*, when unit i is sampled, it takes the value 1 and 0 otherwise. By r_i we denote the 0-1-response indicator for unit i . If unit i is sampled and did respond then $r_i = 1$. It is 0 otherwise. The sample size is n . Finally, π_i denotes the first-order inclusion probability of unit i .

The key to our definitions lies in the individual response propensities. Let ρ_i be the probability that unit i responds when it is sampled.

The interpretation of a response propensity is not straightforward by itself. We follow a model-assisted approach, *i.e.*, the only randomness is in the sample and response indicators. A response probability is a feature of a labelled and identifiable unit, a biased coin that the unit carries in a pocket, so to speak, and is, therefore, inseparable from that unit. With a little effort, however, all concepts can be translated into a model-based context.

First, we give a strong definition.

Definition (strong): A response subset is representative with respect to the sample if the response propensities ρ_i are the same for all units in the population

$$\rho_i = P[r_i = 1 | s_i = 1] = \rho, \quad \forall i, \quad (1)$$

and if the response of a unit is independent of the response of all other units.

If a missing-data mechanism would satisfy the strong definition, then the mechanism would correspond to

Missing-Completely-at-Random (MCAR) with respect to all possible survey questions. Although the definition is appealing, the validity of it can never be tested in practice. We have no replicates of the response of one single unit. We, therefore, also construct a weak definition that can be tested in practice.

Definition (weak): A response subset is representative of a categorical variable X with H categories if the average response propensity over the categories is constant

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \quad \text{for } h = 1, 2, \dots, H, \quad (2)$$

where N_h is the population size of category h , ρ_{hk} is the response propensity of unit k in class h and summation is over all units in this category.

The weak definition corresponds to a missing-data mechanism that is MCAR with respect to X , as MCAR states that we cannot distinguish respondents from non-respondents based on knowledge of X .

3. R-indicators

In the previous section, we defined strong and weak representative response. Both definitions make use of individual response probabilities that are unknown in practice. First, we start with a population R-indicator. From there on, we base the same R-indicator on a sample and on estimated response propensities.

3.1 Population R-indicators

We first consider the hypothetical situation where the individual response propensities are known. Clearly, in that case we can even test the strong definition and we simply want to measure the amount of variation in the response propensities; the more variation, the less representative in the strong sense. Let $\rho = (\rho_1, \rho_2, \dots, \rho_N)'$ be a vector of response propensities, let $\mathbf{1} = (1, 1, \dots, 1)'$ be the N -vector of ones, and let $\rho_0 = \mathbf{1} \times \bar{\rho}$ be the vector consisting of the average population propensity.

Any distance function d in $[0, 1]^N$ would suffice in order to measure the deviation from a strong representative response by calculating $d(\rho, \rho_0)$. Note that the height of the overall response does not play a role. The Euclidean distance is a straightforward distance function. When applied to a distance between ρ and ρ_0 , this measure is proportional to the standard deviation of the response probabilities

$$S(\rho) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\rho_i - \bar{\rho})^2}. \quad (3)$$

It is not difficult to show that

$$S(\rho) \leq \sqrt{\bar{\rho}(1 - \bar{\rho})} \leq \frac{1}{2}. \quad (4)$$

We want the R-indicator to take values on the interval $[0, 1]$ with the value 1 being strong representativeness and the value 0 being the maximum deviation from strong representativeness. We propose the R-indicator R , which is defined by

$$R(\rho) = 1 - 2S(\rho). \quad (5)$$

Note that the minimum value of (5) depends on the response rate, see figure 1. For $\bar{\rho} = 1/2$, it has a minimum value of 0. For $\bar{\rho} = 0$ and $\bar{\rho} = 1$, clearly no variation is possible and the minimum value is 1. Paradoxically, the lower bound increases when the response rate decreases from $1/2$ to 0. For a low response rate, there is less room for individual response propensities to have a large variation.

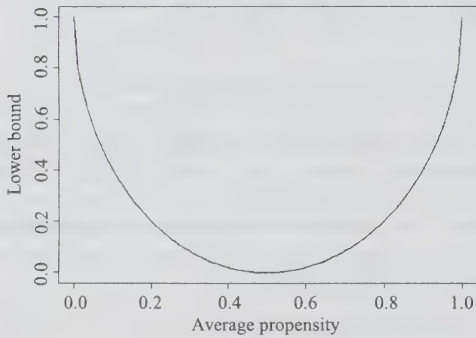


Figure 1 Minimum value of R-indicator (5) as a function of the average response propensity

One may view R as a lack of association measure. When $R(\rho) = 1$ there is no relation between any survey item and the missing-data mechanism. We show that R in fact has a close relation to the well-known χ^2 -statistic that is often used to test independence and goodness-of-fit.

Suppose that the response propensities are only different for classes h defined by a categorical variable X . Let $\bar{\rho}_h$ and f_h be, respectively, the response propensity and the population function of class h , i.e.,

$$f_h = \frac{N_h}{N}, \quad \text{for } h = 1, 2, \dots, H. \quad (6)$$

Hence, for all i with $X_i = h$ the response propensity is $\rho_i = \bar{\rho}_h$.

Since the variance of the response propensities is the sum of the 'between' and 'within' variances over classes h , and the within variances are assumed to be zero, it holds that

$$\begin{aligned} S^2(\bar{\rho}) &= \frac{1}{N-1} \sum_{h=1}^H N_h (\bar{\rho}_h - \bar{\rho})^2 \\ &= \frac{N}{N-1} \sum_{h=1}^H f_h (\bar{\rho}_h - \bar{\rho})^2 \approx \sum_{h=1}^H f_h (\bar{\rho}_h - \bar{\rho})^2. \end{aligned} \quad (7)$$

The χ^2 -statistic measures the distance between observed and expected proportions. However, it is only a true distance function in the mathematical sense for fixed marginal distributions f_h and $\bar{\rho}$. We can apply the χ^2 -statistic to X in order to 'measure' the distance between the true response behaviour and the response behaviour that is expected when response is independent of X . In other words, we measure the deviation from weak representativeness with respect to X .

We can rewrite the χ^2 -statistic to get

$$\begin{aligned} \chi^2 &= \sum_{h=1}^H \frac{(N_h \bar{\rho}_h - N_h \bar{\rho})^2}{N_h \bar{\rho}} \\ &+ \sum_{h=1}^H \frac{(N_h(1 - \bar{\rho}_h) - N_h(1 - \bar{\rho}))^2}{N_h(1 - \bar{\rho})} \\ &= \sum_{h=1}^H \frac{N f_h (\bar{\rho}_h - \bar{\rho})^2}{\bar{\rho}} + \sum_{h=1}^H \frac{N f_h (\bar{\rho}_h - \bar{\rho})^2}{(1 - \bar{\rho})} \\ &= \frac{N}{\bar{\rho}(1 - \bar{\rho})} \sum_{h=1}^H f_h (\bar{\rho}_h - \bar{\rho})^2 \\ &= \frac{N-1}{\bar{\rho}(1 - \bar{\rho})} S^2(\bar{\rho}). \end{aligned} \quad (8)$$

An association measure that transforms the χ^2 -statistic to the $[0, 1]$ interval, see e.g., Agresti (2002), is Cramér's V

$$V = \sqrt{\frac{\chi^2}{N(\min\{C, R\} - 1)}}, \quad (9)$$

where C and R are, respectively, the number of columns and rows in the underlying contingency table. Cramér's V attains a value 0 if observed proportions exactly match expected proportions and its maximum is 1. In our case, the denominator equals N since the response indicator has only two categories: response and non-response. As a consequence, (9) changes into

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N-1}{N\bar{\rho}(1 - \bar{\rho})}} S(\bar{\rho}). \quad (10)$$

From (10) we can see that for large N , Cramér's V is approximately equal to the standard deviation of the response propensities standardized by the maximal standard deviation $\sqrt{\bar{\rho}(1 - \bar{\rho})}$ for a fixed average response propensity $\bar{\rho}$.

3.2 Response-based R-indicators

In section 3.1, we assumed that we know the individual response propensities. Of course, in practice these propensities are unknown. Furthermore, in a survey, we only have information about the response behaviour of sample units. We, therefore, have to find alternatives to the indicators R . An obvious way to do this is to use response-based estimators for the individual response propensities and the average response propensity.

We let $\hat{\rho}_i$ denote an estimator for ρ_i which uses all or a subset of the available auxiliary variables. Methods that support such estimation are, for instance, logistic or probit regression models (Agresti 2002) and CHAID classification trees (Kass 1980). By $\hat{\bar{\rho}}$ we denote the weighted sample average of the estimated response propensities, *i.e.*,

$$\hat{\bar{\rho}} = \frac{1}{N} \sum_{i=1}^N \hat{\rho}_i \frac{s_i}{\pi_i}, \quad (11)$$

where we use the inclusion weights.

We replace R by the estimators \hat{R}

$$\hat{R}(\rho) = 1 - 2 \sqrt{\frac{1}{N-1} \sum_{i=1}^N \frac{s_i}{\pi_i} (\hat{\rho}_i - \hat{\bar{\rho}})^2}. \quad (12)$$

Note that in (12) there are in fact two estimation steps based on different probability mechanisms. The response propensities themselves are estimated and the variation in the propensities is estimated. We return to the consequences of the two estimation steps in section 4.

3.3 Example

We apply the proposed R-indicators to the survey data from the 1998 POLS that we described in section 2.1. Recall that the survey was a combination of face-to-face and telephone interviewing in which the first month was CAPI only. The sample size was close to 40,000 and the response rate was approximately 60%. We linked the fieldwork administration to the sample and deduced whether each contact attempt resulted in a response. This way, we can monitor the pattern of the R-indicator during the fieldwork period.

For the estimation of response rates we used a logistic regression model with region, ethnic background and age as independent variables. Region was a classification with 16 categories, the 12 provinces and the four largest cities – Amsterdam, Rotterdam, The Hague and Utrecht – as separate categories. Ethnic background has seven categories: native, Moroccan, Turkish, Surinam, Dutch Antilles, other non-western non-native and other western non-native. The classification is based on the country of birth of the parents of the selected person. The age variable has three categories: 0 – 34 years, 35 – 54 years, and 55 years and older.

In figure 2, \hat{R} is plotted against the response rate for the first six contact attempts in POLS. The leftmost value corresponds to the respondent pool after one attempt was made. For each additional attempt, the response rate increases but the indicator shows a drop in representativeness. This result confirms the findings in Schouten (2004).

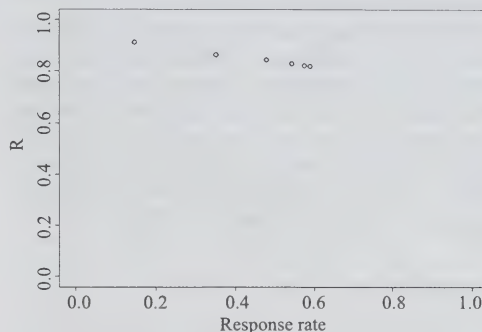


Figure 2 R-indicator for first six contact attempts in POLS 1998

4. Features of R-indicators

In section 3, we propose a candidate indicator for representativeness. However, other indicators can be constructed. There are many association measures or fit indexes, *e.g.*, Goodman and Kruskal (1979), Bentler (1990) and Marsh, Balla and McDonald (1988). Association measures have a strong relation to R-indicators. Essentially, R-indicators attempt to measure in a multivariate setting the lack of association. In this section, we discuss the desired features of R-indicators. We show that the proposed R-indicator R allows for a straightforward upper bound on the non-response bias.

4.1 Features in general

We want R-indicators to be based on a distance function or metric in the mathematical sense. The triangle inequality property of a distance function allows for a partial ordering of the variation in response propensities which enables interpretation. A distance function can easily be derived from any mathematical norm. In section 3, we chose to use the Euclidean norm as this norm is commonly used. The Euclidean norm led us to an R-indicator that uses the standard deviation of response propensities. Other norms, like the supremum norm, would lead us to alternative distance functions. In section 4.3, however, we show that the Euclidean norm based R-indicators have interesting normalization features.

We must make a subtle distinction between R-indicators and distance functions. Distance functions are symmetric while an R-indicator measures a deviation with respect to a specific point, namely the situation where all response propensities are equal. If we change the vector of individual propensities, then this point is in most cases shifted. However, if we fix the average response propensity, then the distance function facilitates interpretation.

Apart from a relation to a distance function, we want to be able to measure, interpret and normalize the R-indicators. In section 3.2, we already derived response-based estimators for 'population' R-indicators that are not measurable when response propensities are unknown and all we have is the response to a survey. Hence, we made R-indicators measurable by switching to estimators. The other two features are discussed separately in the next two sections.

4.2 Interpretation

The second feature of R-indicators is the ease with which we can interpret their values and the concept they are measuring. We moved to an estimator for an R-indicator that is based on the samples of surveys and on estimators of individual response probabilities. Both have far-reaching consequences for the interpretation and comparison of the R-indicator.

Since the R-indicator is an estimator itself, it is also a random variable. This means that it depends on the sample, *i.e.*, it is potentially biased and has a certain accuracy. But what is it estimating?

Let us first assume that the sample size is arbitrarily large so that precision does not play a role and also suppose the selection of a model for response propensities is no issue. In other words, we are able to fit any model for any fixed set of auxiliary variables.

There is a strong relation between the R-indicator and the availability and use of auxiliary variables. In section 2, we defined strong and weak representativeness. Even in the case where we are able to fit any model, we are not able to estimate response propensities beyond the 'resolution' of the available auxiliary variables. Hence, we can only draw conclusions about weak representativeness with respect to the set of auxiliary variables. This implies that whenever an R-indicator is used, it is necessary to complement its value by the set of covariates that served as a grid to estimate individual response propensities. If the R-indicator is used for comparative purposes, then those sets must be the same. We must add that it is not necessary for all auxiliary variables to be used for the estimation of propensities, since they may not add any explanatory power to the model. However, the same sets should be available. The R-indicator then measures a deviation from weak representativeness.

The R-indicator does not capture differences in response probabilities within subgroups of the population other than the subgroups defined by the classes of X . If we let $h = 1, 2, \dots, H$ again denote strata defined by X , N_h be the size of stratum h , and \bar{p}_h be the population average of the response probabilities in stratum h , then it is not difficult to show that \hat{R} is a consistent estimator of

$$R_X(\rho) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{h=1}^H N_h (\bar{p}_h - \bar{p})^2}, \quad (13)$$

when standard models like logistic regression or linear regression are used to estimate the response probabilities. Of course, (13) and (5) may be different.

In practice, the sample size is not arbitrarily large. The sample size affects both estimation steps; the estimation of response propensities and the estimation of the R-indicator using a sample.

If we knew the individual response propensities, then the sample-based estimation of the R-indicator would only lead to variance and not to bias. We would be able to estimate the population R-indicator without bias. Hence, for small sample sizes, the estimators would have a small precision which could be accounted for by using confidence intervals instead of merely point estimators.

The implications for the estimation of response probabilities are, however, different because of model selection and model fit. There are two alternatives. Either one imposes a model to estimate propensities fixing the covariates beforehand, or one lets the model be dependent on the significant contribution of covariates with respect to some predefined level. In the first case, again no bias is introduced but the standard error may be affected by over fitting. In the second case, the model for the estimation of response propensities depends on the size of the sample; the larger the sample, the more interactions that are accepted as significant. Although it is standard statistical practice to fit models based on a significance level, model selection may introduce bias and variance to the estimation of any R-indicator. This can be easily understood by going to the extreme of a sample of, say, size 10. For such a small sample, no interaction between response behaviour and auxiliary characteristics will be accepted, leaving an empty model and an estimated R-indicator of 1. Small samples simply do not allow for the estimation of response propensities. In general, a smaller sample size will, thus, lead to a more optimistic view on representativeness.

We should make a further subtle distinction. It is possible that, for one survey, a lot of interactions contribute to the prediction of response propensities but each one contributes very little, while in another survey there is only one but it strongly contributes a single interaction. None of the small contributions may be significant, but together they are as

strong as the one large contribution that is significant. Hence, we would be more optimistic in the first example even if sample sizes would be comparable.

These observations show that one should always use an R-indicator with some care. It cannot be viewed as separate from the auxiliary variables that were used to compute it. Furthermore, the sample size has an impact on both bias and precision.

4.3 Normalization

The third important feature is the normalization of an R-indicator. We want to be able to attach bounds to an R-indicator so that the scale of an R-indicator, and, hence, changes in the R-indicator get a meaning. Clearly, the interpretation issues that we raised in the previous section also affect the normalization of the R-indicator. Therefore, in this section we assume the ideal situation where we can estimate response propensities without bias. This assumption holds for large surveys. We discuss the normalization of the R-indicator \hat{R} .

4.3.1 Maximal absolute bias and maximal root mean square error

We show that for any survey item Y , the R-indicator can be used to set upper bounds to the non-response bias and to the root mean square error (RMSE) of adjusted response means. We use these bounds of the R-indicator to show the impact under worst-case scenarios.

Let Y be some variable that is measured in a survey and let \hat{y}_{HT} be the Horvitz-Thompson estimator for the population mean based on the survey response. It can be shown (e.g., Bethlehem 1988, Särndal and Lundström 2005) that its bias $B(\hat{y}_{HT})$ is approximately equal to

$$B(\hat{y}_{HT}) = \frac{C(y, \rho)}{\bar{\rho}}, \quad (14)$$

with $C(y, \rho) = 1/N \sum_{i=1}^N (y_i - \bar{y})(\rho_i - \bar{\rho})$ the population covariance between the survey items and the response probabilities. For a close approximation of the variance $s^2(\hat{y}_{HT})$ of \hat{y}_{HT} we refer to Bethlehem (1988).

A normalization of R is found by the Cauchy-Schwarz inequality. This inequality states that the covariance between any two variables is bounded in absolute sense by the product of the standard deviations of the two variables. We can translate this to bounds for the bias (14) of \hat{y}_{HT}

$$\begin{aligned} |B(\hat{y}_{HT})| &\leq \frac{S(\rho)S(y)}{\bar{\rho}} = \frac{(1 - R(\rho))S(y)}{2\bar{\rho}} \\ &= B_m(\rho, y). \end{aligned} \quad (15)$$

Clearly, we do not know the upper bound $B_m(\rho, y)$ in (15) but we can estimate it using the sample and the estimated response probabilities. We denote the estimator by $\hat{B}_m(\hat{\rho}, y)$.

In a similar way, we can set a bound to the root mean square error (RMSE) of \hat{y}_{HT} . It holds approximately that

$$\begin{aligned} \text{RMSE}(\hat{y}_{HT}) &= \sqrt{B^2(\hat{y}_{HT}) + s^2(\hat{y}_{HT})} \\ &\leq \sqrt{B_m^2(\rho, y) + s^2(\hat{y}_{HT})} \\ &= E_m(\rho, y). \end{aligned} \quad (16)$$

Again, we do not know $E_m(\rho, y)$. Instead, we use the sample-based estimator that employs the estimated response probabilities, denoted by $\hat{E}_m(\hat{\rho}, y)$.

The bounds $\hat{B}_m(\hat{\rho}, y)$ and $\hat{E}_m(\hat{\rho}, y)$ are different for each survey item y . For comparison purposes it is, therefore, convenient to define a hypothetical survey item. We suppose that $\hat{S}(y) = 0.5$. The corresponding bounds we denote by $\hat{B}_m(\hat{\rho})$ and $\hat{E}_m(\hat{\rho})$. They are equal to

$$\hat{B}_m(\hat{\rho}) = \frac{(1 - \hat{R}(\hat{\rho}))}{4\hat{\rho}} \quad (17)$$

$$\hat{E}_m(\hat{\rho}) = \sqrt{\hat{B}_m^2(\hat{\rho}) + \hat{s}^2(\hat{y}_{HT})}. \quad (18)$$

We compute (17) and (18) in all studies described in section 5. We have to note that (17) and (18) are again random variables that have a certain precision and that are potentially biased.

4.3.2 Response-representativeness functions

In the previous section, we used the R-indicator to set upper bounds to the non-response bias and to the root mean square error of the (adjusted) response mean. Conversely, we may set a lower bound to the R-indicator by demanding that either the absolute non-response bias or the root mean square error is smaller than some prescribed value. Such a lower bound may be chosen as one of the ingredients of quality restrictions put upon the survey data by a user of the survey. If a user does not want the non-response bias or root mean square to exceed a certain value, then the R-indicator must be bigger than the corresponding bound.

Clearly, lower bounds to the R-indicator depend on the survey item. Therefore, again we restrict ourselves a hypothetical survey item for which $\hat{S}(y) = 0.5$.

It is not difficult to show from (17) that if we demand that

$$\hat{B}_m(\hat{\rho}) \leq \gamma, \quad (19)$$

then it must hold that

$$\hat{R} \geq 1 - 4\hat{\rho}\gamma = r_1(\gamma, \hat{\rho}). \quad (20)$$

Analogously, using (18) and demanding that

$$\hat{E}_m(\hat{\rho}) \leq \gamma, \quad (21)$$

we arrive at

$$\hat{R} \geq 1 - 4\hat{\rho}\sqrt{\gamma^2 - \hat{s}^2(\hat{y}_{HT})} = r_2(\gamma, \hat{\rho}). \quad (22)$$

In (20) and (22) we let $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ denote lower limits to the R-indicator. In the following section, we refer to $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ as response-representativeness functions. We compute them for the studies in section 5.

4.3.3 Example

We again illustrate the normalization with the same example used in sections 2.1 and 3.3. Figure 3 contains the response-representativeness function $r_1(\gamma, \hat{\rho})$ and the observed R-indicators \hat{R} for the six contact attempts in POLS 1998. Three values of γ are chosen, $\gamma = 0.1$; $\gamma = 0.075$ and $\gamma = 0.05$.

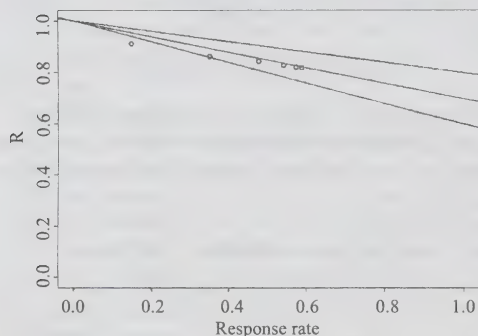


Figure 3 Lower bounds for R-indicator \hat{R} for the first six contact attempts of POLS 1998. Lower bounds are based on $\gamma = 0.1$, $\gamma = 0.075$ and $\gamma = 0.05$

Figure 3 indicates that after the second contact attempt, the values of the R-indicator exceed the lower bound corresponding to the 10%-level. After four attempts, the R-indicator is close to the 7.5%-level. However, the values never exceed the other lower bound that is based on the 5%-level.

In figure 4, the maximal absolute bias $\hat{B}_m(\hat{\rho})$ is plotted against the response rate of the six contact attempts. After the third contact attempt, the R-indicator has converged on a value around 8%.

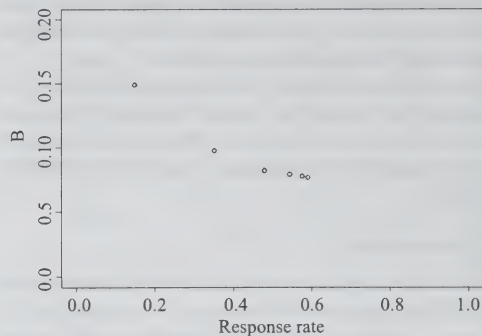


Figure 4 Maximal absolute bias for the first six contact attempts of POLS 1998

5. Application of the R-indicator

In this section, we apply the R-indicator to two studies that investigate different non-response follow-up strategies and different combinations of data collection modes. The first study involves the Dutch Labour Force Survey (LFS). The study is an investigation of both the call-back approach (Hansen and Hurwitz 1946) and the basic-question approach (Kersten and Bethlehem 1984). The second study deals with mixed-mode data collection designs applied to the Dutch Safety Monitor survey.

In sections 5.2 and 5.3 we take a closer look at the studies in connection with the representativeness of their different fieldwork strategies. First, in section 5.1 we describe how we approximate standard errors.

5.1 Standard error and confidence interval

If we want to compare the values of the R-indicator for different surveys or data collection strategies, we need to estimate their standard errors.

The R-indicator \hat{R} involves the sample standard deviation of the estimated response probabilities. This means that there are two random processes involved. The first process is the sampling of the population. The second process is the response mechanism of the sampled units. If the true response probabilities were known, then drawing a sample would still introduce uncertainty about the population R-indicator and, hence, lead to a certain loss of precision. However, since we do not know the true response probabilities, these probabilities are estimated using the sample. This introduces additional precision loss.

An analytical derivation of the standard error of \hat{R} is not straightforward due to the estimation of the response probabilities. In this paper, we are resigned to naïve numerical

approximations of the standard error. We estimate the standard error of the R-indicator by non-parametric bootstrapping (Efron and Tibshirani 1993). The non-parametric bootstrap estimates the standard error of the R-indicator by drawing a number $b = 1, 2, \dots, B$ of so-called bootstrap samples. These are samples drawn independently and with replacement from the original dataset, of the same size n as the original dataset. The R-indicator is calculated for every bootstrap sample b . We thus obtain B replications of the R-indicator; \hat{R}_b^{BT} , $b = 1, 2, \dots, B$. The standard error for the empirical distribution of these B replications is an estimate for the standard error of the R-indicator, that is

$$s_R^{BT} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{R}_b^{BT} - \hat{\bar{R}}^{BT})^2} \tag{23}$$

where $\hat{\bar{R}}^{BT} = 1/B \sum_{b=1}^B \hat{R}_b^{BT}$ is the average estimated R-indicator.

In the approximations, we take $B = 200$ for all studies. We experimented with larger numbers of B of up to $B = 500$, but found that in all cases, the estimate of the standard error had converged by $B = 200$.

We determine $100(1 - \alpha)\%$ confidence intervals by assuming a normal approximation of the distribution of \hat{R} employing the estimated standard errors using (23)

$$CI_{\alpha}^{BT} = (\hat{R} \pm \xi_{1-\alpha} \times s_R^{BT}) \tag{24}$$

with $\xi_{1-\alpha}$ the $1 - \alpha$ quantile of the standard normal distribution.

5.2 Labour Force Survey; follow-up study 2005

From July to December 2005, Statistics Netherlands conducted a large-scale follow-up of non-respondents in the Dutch Labour Force Survey (LFS). In the study, two samples of non-respondents in the LFS were approached once more using either a call-back approach (Hansen and Hurwitz 1946) or a basic-question approach (Kersten and Bethlehem 1984). The samples consisted of LFS households that refused, were not processed or were not contacted in the LFS for the months July–October. In the design of the follow-up study, we used the recommendations in the studies by Stoop (2005) and Voogt (2004).

The main characteristics of the call-back and basic-question approaches applied to the LFS are given in Table 2. For more details, we refer to Schouten (2007) and Cobben and Schouten (2007). The call-back approach employed the original household questionnaire in CAPI, while the basic-question approach used short questionnaires in a mixed-mode setting. The mixed-mode design involved web, paper and CATI. CATI was used for all households with a listed phone number. Households without a listed phone number received an advance letter, a paper questionnaire and a login

to a secure website containing the web questionnaire. Respondents were left the choice to fill in either the paper or web questionnaire.

Table 2
Characteristics of the two approaches in the follow-up study

Call-back approach	Basic-question approach
<ul style="list-style-type: none">• LFS questionnaire to be answered by all members of the household in CAPI• 28 interviewers geographically selected from historically best-performing interviewers• Interviewer was different from interviewer that received non-response• Interviewers received additional training in doorstep interaction• Extended fieldwork period of two months• Interviewer could offer incentives• Interviewers could receive a bonus• A paper summary of the characteristics of the non-responding household was sent to the interviewer• Allocation of address one week after non-response	<ul style="list-style-type: none">• A strongly condensed questionnaire with key questions of the LFS which takes between 1 and 3 minutes to answer or fill in• Mixed-mode data collection design using web, paper and CATI• The questionnaire was to be answered by one person per household following the next birthday method• The timing is one week after the household is processed as a non-response

The sample size of the LFS pilot was $n = 18,074$ households, of which 11,275 households responded. The non-responding households were stratified according to the cause of non-response. Households that were not processed or contacted, and households that refused were eligible for a follow-up. It was considered to be unethical to follow-up households that did not respond due to other causes like illness. In total, 6,171 households were eligible. From these households, two simple random samples were drawn of size 775. In the analyses, the non-sampled eligible households were left out. The sampled eligible households received a weight accordingly. The 11,275 LFS respondents and the 628 ineligible households all received a weight of one. This implies that the inclusion probabilities are unequal for this example.

Schouten (2007) compared the LFS respondents to the converted and persistent non-respondents in the call-back approach using a large set of demographic and socio-economic characteristics. He used logistic regression models to predict the type of response. He concluded that the

converted non-respondents in the call-back approach are different from the LFS respondents with respect to the selected auxiliary variables. Furthermore, he found no evidence that the converted non-respondents were different from persistent non-respondents with respect to the same characteristics. These findings have led to the conclusion that the combined response of the LFS and call-back approach is more representative with respect to the selected auxiliary variables.

The additional response in the basic question approach was analyzed by Cobben and Schouten (2007) using the same set of auxiliary variables and employing the same logistic regression models. For this follow-up, the findings were different for households with and without a listed phone number. When restricted to listed households, they found the same results as for the call-back approach; the response becomes more representative after the addition of the listed basic-question respondents. However, for the overall population, *i.e.*, including the unlisted households, the inverse was found. The basic-question approach gives 'more of the same' and, hence, sharpens the contrast between respondents and non-respondents. Combining LFS response with basic-question response leads to a less representative composition. In the logistic regression models by Cobben and Schouten (2007) the 0-1 indicators for having a listed phone number and having a paid job gave a significant contribution.

Cobben and Schouten (2007) and Schouten (2007) used the set of auxiliary variables listed in Table 3. The auxiliary variables were linked to the sample from various registers and administrative data. The variables in logistic regression models for response probabilities were selected when the variables gave a significant contribution at the 5% level. Otherwise, they were excluded.

Table 3

The auxiliary variables in the studies by Schouten (2007) and Cobben and Schouten (2007). The household core is the head of the household and his or her partner if present

Variable
Household has a listed phone number
Region of the country in 4 classes
Province and 4 largest cities
Average age in 6 classes
Ethnic group in 4 classes
Degree of urbanization in 5 classes
Household type in 6 classes
Gender
Average house value at zip code level in 11 classes
At least one member of household core is self-employed
At least one member of household core has a subscription to the CWI
At least one member of household core receives social allowance
At least one member of household core has a paid job
At least one member of household core receives disability allowance

Table 4 shows the weighted sample size, response rate, \hat{R} , $CI_{0.05}^{BT}$, \hat{B}_m and \hat{E}_m for the response to the LFS, the response of the LFS combined with the call-back response and the response of the LFS combined with the basic-question response. The standard errors are relatively large with respect to the studies in subsequent sections due to the weighting. There is an increase in \hat{R} when the call-back respondents are added to the LFS respondents. As both the response rate and the R-indicator increase, the maximal absolute bias \hat{B}_m decreases. The confidence intervals $CI_{0.05}^{BT}$ for the LFS response and the combined LFS and call-back response overlap. However, the one-sided null hypothesis $H_0: R_{LFS} - R_{LFS+CB} \geq 0$ is rejected at the 5%-level.

Table 4

Weighted sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, LFS plus call-back, and LFS plus basic-question for the extended set of auxiliary variables

Response	n	Rate	\hat{R}	$CI_{0.05}^{BT}$	\hat{B}_m	\hat{E}_m
LFS	18,074	62.2%	80.1%	(77.5-82.7)	8.0%	8.0%
LFS + call-back	18,074	76.9%	85.1%	(82.4-87.8)	4.8%	4.9%
LFS + basic-question	18,074	75.6%	78.0%	(75.6-80.4)	7.3%	7.3%

In Table 4, there is a decrease in \hat{R} when we compare the LFS response to the combined response with the basic-question approach. This decrease is not significant. \hat{B}_m slightly decreases. In Table 5, this comparison is restricted to households with a listed phone number. The R-indicator in general is much higher than for all the households. Because the sample size is now smaller, the estimated standard errors are larger as is reflected in the width of the confidence interval. \hat{B}_m is decreased. For the combined response in the LFS and the basic-question approach, we see an increase of \hat{R} but again this increase is not significant. \hat{B}_m decreases.

Table 5

Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, and LFS plus basic-question restricted to households with listed phone numbers and for the extended set of auxiliary variables

Response	n	Rate	\hat{R}	$CI_{0.05}^{BT}$	\hat{B}_m	\hat{E}_m
LFS	10,135	68.5%	86.3%	(83.1-89.5)	5.0%	5.1%
LFS + basic-question	10,135	83.0%	87.5%	(84.3-90.7)	3.8%	3.8%

We find in the example of the LFS follow-up that the R-indicators confirm the conclusions for the call-back approach and the basic question approach. Furthermore, the increase in the R-indicator that follows by adding the call-back response is significant at the 5% level.

5.3 Safety Monitor; pilot mixed-mode 2006

In 2006, Statistics Netherlands conducted a pilot on the Safety Monitor to investigate mixed-mode data collection strategies. See Cobben, Janssen, Berkel and Brakel (2007) for details. The regular Safety Monitor surveys individuals of 15 years and older in the Netherlands about issues that relate to safety and police performance. The Safety Monitor is a mixed-mode survey. Persons with a listed phone number are approached by CATI. Persons that cannot be reached by telephone are approached by CAPI. In the 2006 pilot, the possibility of using the Internet as one of the modes in a mixed-mode strategy was evaluated. Persons in the pilot were first approached with a web survey. Non-respondents to the web survey were re-approached by CATI when they had a listed phone number and by CAPI otherwise. In Table 6 we give the response rates for the normal survey, the pilot response to the web only, and the response to the pilot as a whole. The response to the web survey alone is low. Only 30% of the persons filled in the web questionnaire. This implied that close to 70% of the sampled units were re-allocated to either CAPI or CATI. This resulted in an additional response of approximately 35%. The overall response rate is slightly lower than that of the normal survey.

Fouwels, Janssen and Wetzels (2006) performed a univariate analysis of response compositions. They argue that the response rate is lower for the pilot but that this decrease is quite stable over various demographic sub-groups. They observe a univariate decline in response rate for the auxiliary variables age, ethnic group, degree of urbanization and type of household. However, they do find indications that the response becomes less representative when the comparison is restricted to the web respondents only. This holds, not surprisingly, especially for the age of the sampled persons.

Table 6 contains the sample size, response rate, \hat{R} , $CI_{0.05}^{BT}$, \hat{B}_m and \hat{E}_m for three groups: the regular survey, the pilot survey restricted to web and the pilot survey as a whole. The auxiliary variables age, ethnic group, degree of urbanization and type of household were linked from registers and were selected in the logistic model for the response probabilities. Table 6 shows that the R-indicator for the web response is lower than that of the regular survey. The corresponding p -value is close to 5%. As a consequence of both a low response rate and a low R-indicator, the maximal absolute bias \hat{B}_m is more than twice as high as for the regular survey. However, for the pilot as a whole, both the R-indicator and \hat{B}_m are close to the values of the regular survey. Due to the smaller sample size of the pilot, the estimated standard errors are larger than in the regular survey.

Table 6

Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response for the regular Safety Monitor, the pilot with web only and the pilot with web and CAPI/CATI follow-up

Response	n	Rate	\hat{R}	$CI_{0.05}^{BT}$	\hat{B}_m	\hat{E}_m
Regular	30,139	68.9%	81.4%	(80.3-82.4)	6.8%	6.8%
Pilot - web	3,615	30.2%	77.8%	(75.1-80.5)	18.3%	18.4%
Pilot - web plus	3,615	64.7%	81.2%	(78.3-84.0)	7.3%	7.4%

The findings in Table 6 do not contradict those of Fouwels *et al.* (2006). We also find that the web response in the pilot has a less balanced composition, whereas the composition of the full pilot response is not markedly worse than that of the Safety Monitor itself.

6. Discussion

We have three main objectives in this paper: a mathematically rigorous definition and perception of representative response, the construction of a potential indicator for representativeness, and the empirical illustrations of the indicator's use. As we saw, the proposed indicator is an example of what we call R-indicators, where 'R' stands for representativeness. With the empirical illustration, we want to find support for the idea that such R-indicators are valuable tools in the comparison of different surveys and data collection strategies. R-indicators are useful if they confirm findings in elaborate analyses of studies that involve multiple surveys in time or on a topic.

The R-indicator in this paper is promising because it can easily be computed and allows for interpretation and normalization when response propensities can be estimated without error. The application to real survey data shows that the R-indicator confirms earlier analyses of the non-response composition. Other R-indicators can, of course, simply be constructed by choosing different distance functions between vectors of response propensities. The R-indicator and graphical displays showed in this paper can be computed using most standard statistical software packages.

The computation of R-indicators is sample-based and employs models for individual response propensities. Hence, R-indicators are random variables themselves and there are two estimation steps that influence their bias and variance. However, it is mostly the modelling of response propensities that has important implications. The restriction to the sample for the estimation of R-indicators implies that those indicators are less precise, but this restriction does not introduce a bias asymptotically. Model selection and model fit usually are performed by choosing a significance level and adding only those interactions to the model that give a significant contribution. The latter means that the size of the

sample and the availability of auxiliary variables play an important role in the estimation of response propensities. Bias may be introduced by the model selection strategy. There are various obvious approaches for dealing with the dependence on the size of the sample. One may not do a model selection but fix a stratification beforehand. That way, bias is avoided but standard errors are not controlled and may be considerable. One may also let empirical validation be the input to develop 'best practices' for R-indicators.

We applied the proposed R-indicator to two studies that were conducted at Statistics Netherlands in recent years, and that were thoroughly investigated by other authors. The increase or decrease in the R-indicator conforms to the more detailed analyses done by these authors. We, therefore, conclude that R-indicators can be valuable tools. However, more empirical evidence is clearly needed.

The application of the R-indicator showed that there is no clear relation between response rate and representativeness of response. Larger response rates do not necessarily lead to a more balanced response. Not surprisingly, we do find that higher response rates reduce the risk of non-response bias. The higher the response rate, the smaller the maximal absolute bias of survey items.

Application to the selected studies showed that standard errors do decrease with increasing sample size as expected, but they are still relatively large for modest sample sizes. For example, for a sample size of 3,600, we found a standard error of approximately 1.3%. Hence, if we assume a normal distribution, then the 95% confidence interval has an approximate width of 5.4%. The sample size of the LFS is about 30,000 units. The standard error is approximately 0.5% and the corresponding 95% confidence interval is approximately 2% wide. The standard errors are larger than we expected.

This paper contains a first empirical study of an R-indicator and its standard error. Much more theoretical and empirical research is necessary to fully understand R-indicators and their properties. First, we did not consider survey items at all. Clearly, it is imperative that we do this in the future. However, as we already argued, R-indicators are dependent on the set of auxiliary variables. It can, therefore, be conjectured that, as for non-response adjustment methods, the extent to which R-indicators predict non-response bias of survey items is dependent on the missing-data mechanism. In a missing-data mechanism that is strongly non-ignorable, R-indicators will not do a good job. However, without knowledge about the missing-data mechanism, no other indicator would either. For this reason, we constructed the notion of maximal absolute bias, as this gives a limit to non-response bias under the worst-case scenario. A second topic of future research is a theoretical derivation of the standard error of the R-indicator used in

this paper. The non-parametric bootstrap errors only give naïve approximations. However, if we want R-indicators to play a more active role in the comparison of different strategies, then we need (approximate) closed forms. Third, we will need to investigate the relation between the selection and number of auxiliary variables and the standard errors of the R-indicator.

Acknowledgements

The authors would like to thank Bob Groves, Björn Janssen, Geert Loosveldt, and the associate editor and two referees for their useful comments and suggestions.

References

- Agresti, A. (2002). *Categorical data analysis*. *Wiley Series in Probability and Statistics*. New York: John Wiley & Sons, Inc., NY, USA.
- Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 2, 238-246.
- Bertino, S. (2006). A measure of representativeness of a sample for inferential purposes. *International Statistical Review*, 74, 149-159.
- Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 3, 251-260.
- Cobben, F., Janssen, B., Berkel, K. van and Brakel, J. van den (2007). Statistical inference in a mixed-mode data collection setting. Paper presented at ISI 2007, August 23-29, 2007, Lisbon, Portugal.
- Cobben, F., and Schouten, B. (2007). An empirical validation of R-indicators. Discussion paper, CBS, Voorburg.
- Curtin, R., Presser, S. and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-428.
- Efron, B., and Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- Fouwels, S., Janssen, B. and Wetzels, W. (2006). Experiment mixed-mode waarneming bij de VMR. Technical paper SOO-2007-H53, CBS, Heerlen.
- Goodman, L.A., and Kruskal, W.H. (1979). Measures of association for cross-classifications. Springer-Verlag, Berlin, Duitsland.
- Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 5, 646-675.
- Groves, R.M., and Peytcheva, E. (2006). The impact of nonresponse rates on nonresponse bias: A meta-analysis. Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.
- Groves, R.M., Presser, S. and Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68, 2-31.

- Hájek, J. (1981). Sampling from finite populations. New York: Marcel Dekker, USA.
- Hansen, M.H., and Hurwitz, W.H. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- Heerwegh, D., Abts, K. and Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 1, 3-10.
- Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 2, 119-127.
- Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.
- Kesten, H.M.P., and Bethlehem, J.G. (1984). Exploring an reducing the nonresponse bias by asking the basic question. *Statistical Journal of the United Nations*, ECE 2, 369-380.
- Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.
- Kruskal, W., and Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *International Statistical Review*, 47, 13-24.
- Kruskal, W., and Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review*, 47, 111-123.
- Kruskal, W., and Mosteller, F. (1979c). Representative sampling III: Current statistical literature. *International Statistical Review*, 47, 245-265.
- Little, R.J.A., and Rubin, D.B. (2002). Statistical analysis with missing data. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc., NY, USA.
- Marsh, H.W., Balla, J.R. and McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 3, 391-410.
- Merkle, D.M., and Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A Little). New York: John Wiley & Sons, Inc., 243-258.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Särndal, C., and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Wiley Series in Survey Methodology, John Wiley & Sons, Chichester, England.
- Särndal, C., Swensson, B. and Wretman, J. (2003). Model-assisted survey sampling. Springer Series in Statistics, Springer, New York.
- Schouten, B. (2004). Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998. Discussion paper 04001, CBS, Voorburg, available at website <http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2004/default.htm>.
- Schouten, B., and Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes, Discussion paper 07002, CBS, Voorburg. Available at website <http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2007/default.htm>.
- Stoop, I. (2005). Surveying nonrespondents. *Field Methods*, 16, 23-54.
- Voogt, R. (2004). I am not interested: Nonresponse bias, response bias and stimulus effects in election research. PhD dissertation, University of Amsterdam, Amsterdam.

Stratified balanced sampling

Guillaume Chauvet ¹

Abstract

In the selection of a sample, a current practice is to define a sampling design stratified on subpopulations. This reduces the variance of the Horvitz-Thompson estimator in comparison with direct sampling if the strata are highly homogeneous with respect to the variable of interest. If auxiliary variables are available for each individual, sampling can be improved through balanced sampling within each stratum, and the Horvitz-Thompson estimator will be more precise if the auxiliary variables are strongly correlated with the variable of interest. However, if the sample allocation is small in some strata, balanced sampling will be only very approximate. In this paper, we propose a method of selecting a sample that is balanced across the entire population while maintaining a fixed allocation within each stratum. We show that in the important special case of size-2 sampling in each stratum, the precision of the Horvitz-Thompson estimator is improved if the variable of interest is well explained by balancing variables over the entire population. An application to rotational sampling is also presented.

Key Words: Rotational sampling; Maximum entropy; Cube method; Stratification; Unequal probability sampling.

1. Introduction

In the case of stratified sampling, a population U is partitioned into H subpopulations U_h , $h = 1, \dots, H$ called strata, in which samples S_h , $h = 1, \dots, H$ are selected according to independent sampling designs p_h , $h = 1, \dots, H$, respectively. The inclusion probability of unit k is the probability π_k that unit k is in the sample, and the joint inclusion probability is the probability π_{kl} that two distinct units k and l are jointly in the sample. We will write $\pi = (\pi_k)_{k \in U}$ and $\pi^h = (\pi_k)_{k \in U_h}$. We assume that within each stratum U_h , design $p_h(\cdot)$ is of fixed size. In particular, then, we have $\sum_{k \in U_h} \pi_k = n_h$, $h = 1, \dots, H$, where n_h denotes the allocation in stratum U_h . In the rest of the paper, we assume that all sample sizes for stratum n_h are integers.

The Horvitz-Thompson estimator $\hat{t}_{z\pi} = \sum_{k \in S} z_k / \pi_k = \sum_{h=1}^H \hat{t}_{z\pi}^h$, where $\hat{t}_{z\pi}^h = \sum_{k \in S_h} z_k / \pi_k$, provides an unbiased estimate of $t_z = \sum_{k=1}^H t_z^h$, where $t_z^h = \sum_{k \in U_h} z_k$ denotes the total of the variable (vector) z over U_h . In the particular case where $z_k = y_k$ is scalar, the variance of the Horvitz-Thompson estimator is given by the Sen-Yates-Grundy variance formula:

$$\text{Var}(\hat{t}_{y\pi}) = \frac{1}{2} \sum_{h=1}^H \sum_{k \neq l \in U_h} (\pi_k \pi_l - \pi_{kl}) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2. \quad (1)$$

This variance is small if the strata are homogeneous with respect to the variable of interest, specifically if y_k / π_k is approximately constant within each stratum.

If a vector $\mathbf{x} = (x_1, \dots, x_q)$ of q auxiliary variables is available prior to sample selection for each individual in the population, the sampling within each stratum can be improved with the cube algorithm (Deville and Tillé 2004),

which selects balanced samples. Sampling design $p_h(\cdot)$ is said to be balanced on the \mathbf{x} variables if the equations

$$\hat{t}_{\mathbf{x}\pi}^h = t_{\mathbf{x}}^h \quad (2)$$

are exactly satisfied. The variance of the Horvitz-Thompson estimator is therefore zero for the estimate of the total of the balancing variables. In the particular case where $\mathbf{x} = \pi$, i.e., if the inclusion probability is the only balancing variable, (2) reduces to

$$\sum_{k \in S_h} 1 = \sum_{k \in U_h} \pi_k = n_h. \quad (3)$$

Hence, stratified sampling of fixed size in each stratum is a particular case of balanced sampling. For any given number of constraints, an exactly balanced sample generally cannot be found. Suppose, for example, that population U_h contains 100 individuals on whom is defined a variable x with two possible values, 0 and 1, and that 53 individuals in the population have the value 0 for that variable. Selecting a size-10 equal-probability sample balanced on variable x would mean selecting a sample containing 5,3 individuals for whom $x = 0$ and 4,7 individuals for whom $x = 1$, which is impossible. Consequently, the goal is generally to select an approximately balanced sample, such that

$$\hat{t}_{\mathbf{x}\pi}^h \approx t_{\mathbf{x}}^h. \quad (4)$$

With the cube method (Deville and Tillé 2004), we can select approximately balanced samples on any number of variables, maintaining exactly a predetermined set of inclusion probabilities π . The method is composed of two phases: the flight phase and the landing phase. At each step in the flight phase, we decide at random to either select or permanently discard one of the population units. At the end of the flight phase, we have, in each stratum U_h , a vector

$\pi^{h*} = (\pi_k^*)_{k \in U_h} \in [0, 1]^N$ that satisfies the following conditions:

$$E(\pi^{h*}) = \pi^h, \quad (5)$$

$$\sum_{k \in U_h} \frac{x_k}{\pi_k} \pi_k^* = \sum_{k \in U_h} x_k, \quad (6)$$

$$\text{Card}\{k \in U_h; 0 < \pi_k^* < 1\} \leq q, \quad (7)$$

where E denotes the expectation for the sampling method used in the flight phase. The vector π^{h*} gives the outcome of the flight phase: π_k^* is 1 if unit k is selected, 0 if it is rejected, and between 0 and 1 only if the decision has not been made for unit k after the flight phase. Equations (5) and (6) ensure that the inclusion probabilities and balancing constraints are maintained perfectly at the end of the flight phase. Equation (7) ensures that a decision remains to be made for no more than q individuals in each stratum U_h , where q is the number of balancing variables. The flight phase ends when the balancing constraints can no longer be exactly satisfied. The landing phase consists in defining, conditionally on the outcome of the flight phase, an optimal sampling design defined on the remaining population V . This design is optimal in that it makes it possible to complete the sampling while minimizing the variance, conditionally on the outcome of the flight phase, of the Horvitz-Thompson estimator of the balancing variables. The remaining units are sampled, conditionally on the outcome of the flight phase, with inclusion probabilities $(\pi_k^*)_{k \in V}$, so that the units' unconditional inclusion probabilities $(\pi_k)_{k \in V}$ are maintained exactly.

The measure of entropy associated with a sampling design $p(\cdot)$ defined on population U is given by

$$I(p) = - \sum_{s \subset U} p(s) \log(p(s)),$$

with the convention $0 \log(0) = 0$. Deville and Tillé (2005) have shown that the balanced design with maximum entropy compared with other sampling designs balanced on the same variables and with the same inclusion probabilities can be regarded as the conditional of a Poisson design. Assuming the asymptotic normality of a multivariate Horvitz-Thompson estimator in the case of a Poisson design, they derived a variance approximation formula for the Horvitz-Thompson estimator for a balanced sampling design. In the case of stratified balanced sampling, we have

$$\text{Var}(\hat{t}_{y\pi}) = \sum_{h=1}^H \sum_{k \in U_h} \frac{b_k}{\pi_k^2} (y_k - \beta_h \mathbf{x}_k)^2 \quad (8)$$

where $\beta_h = (\sum_{l \in U_h} b_l \mathbf{x}_l / \pi_l \mathbf{x}_l' / \pi_l)^{-1} \sum_{l \in U_h} b_l \mathbf{x}_l / \pi_l y_l / \pi_l$. Deville and Tillé (2005) offer several approximations for

the b_k . The simplest is $b_k = \pi_k(1 - \pi_k)$. The variance of the Horvitz-Thompson estimator will be small if, in each stratum, variable of interest y is well explained by balancing variables \mathbf{x} .

Sampling will be balanced in each stratum if the number of balancing variables remains small relative to the sample size. In some cases, however, the allocation to each stratum is too small for balanced sampling: if the stratification of the population is very granular, a current practice is to select a size-2 sample in each stratum. In that case, the only condition that can be imposed is a fixed sample size in each stratum.

In the next section, we propose an algorithm based on the cube method that ensures balanced sampling across the entire population for selected variables and exactly maintains the desired allocation within each stratum. Hence, the samples are no longer selected independently in each stratum. Precision is improved in comparison with stratified sampling with fixed sample size in each stratum if the balancing variables are strongly correlated with the variable of interest across the entire population. The algorithm also has the advantage of ensuring approximate balancing in each stratum, and the larger the sample size allocated to the stratum, the more balanced the sampling will be.

2. Stratified balanced sampling with pooling of landing phases

If sample S is selected from U in accordance with the stratified balanced sampling procedure described in section 1, sampling will be balanced in each stratum as long as the landing phase affects a small number of individuals relative to the sample size. Specifically, equation (7) shows that the number of balancing variables must be small relative to the sample allocation in each stratum. In some cases, that constraint cannot be satisfied. The population is often partitioned into very small groups to make the results more relevant, which means decreasing the sample selected in each stratum; the limit generally used is a size-2 sample, which produces an unbiased variance estimator.

Again, we take the case of a population U divided into H strata U_1, \dots, U_H , for which a vector $\mathbf{x}_k = (\pi_k, \mathbf{z}_k')'$ of auxiliary variables is known. We assume that the variable π_k is one of the balancing constraints, to ensure fixed-size sampling. Where the allocation to each stratum is too small for balanced sampling to apply constraints other than fixed size in each stratum, algorithm 1 provides an alternative sampling method. A flight phase is carried out independently in each of the H strata: we write $\pi^{h*} = (\pi_k^*)_{k \in U_h}$, $h = 1, \dots, H$ for the probability vectors obtained at the end of those flight phases, $\pi^* = (\pi_k^*)_{k \in V}$, where V denotes the units that have not yet been sampled or rejected,

and $\mathbf{x}_k^* = (\pi_k^* \mathbf{1}_{k \in U_1}, \dots, \pi_k^* \mathbf{1}_{k \in U_H}, \mathbf{z}_k' \pi_k^* / \pi_k)'$. The probability vector obtained after a final flight phase over the set of remaining units is written $\pi^{**} = (\pi_k^{**})_{k \in V}$. The set of units in stratum U_h that have not yet been sampled or rejected at the end of this new flight phase is denoted W_h .

Algorithm 1: Stratified balanced sampling with pooling of landing phases

- Step 1. Carry out a flight phase, with balancing variables \mathbf{x}_k and inclusion probabilities π_k , independently in each stratum U_h .
- Step 2. Carry out a flight phase, with balancing variables \mathbf{x}_k^* and inclusion probabilities π_k^* , on the set V of units remaining at the end of step 1.
- Step 3. Select a fixed-size sample from each subpopulation W_h , with inclusion probabilities π_k^{**} .

The algorithm is based on a method used by the Institut National de la Statistique et des Études Économiques (INSEE) to select the primary units of the 1999 Master Sample. The Master Sample is a sample of dwellings selected in the 1999 Census for use as a sample frame for household surveys. A detailed description of the sampling design for the Master Sample is provided in Bourdalle, Christine and Wilms (2000). The dwellings are first grouped into urban units and rural units. In the subpopulation of units with fewer than 100,000 residents, a sample of about 6% is selected. We have four auxiliary variables (taxable net income and three age groups). The expected number of sample units is too small for stratified sampling by region, with balanced sampling on the four variables in each region. The regions were therefore grouped into eight super-regions, and the sampling processes were coordinated in such a way as to ensure both overall balanced sampling for the four auxiliary variables in each super-region and a fixed sample size in each region.

A similar method was proposed by Rousseau and Tardieu (2004) for the selection of balanced samples from large frames using the CUBE macro available on INSEE's Web site. The macro's run time is approximately proportional to the square of the population size. Note that Chauvet and Tillé (2006) proposed a fast method of balanced sampling whose run time depends only on the size of the population and which can select balanced samples directly from very large populations. The algorithm was programmed into an SAS macro (see Chauvet and Tillé, 2005) and is also available in the R Sampling Package prepared by Matei and Tillé (2006). In both programs, the second flight phase is performed by adding a constraint associated with each stratum to balancing variables \mathbf{x}_k^* and maintaining the fixed-size condition in each stratum.

Using inclusion probabilities vector π^* conditionally on the outcome of step 1 ensures that inclusion probabilities

vector π is maintained by deconditioning from the outcome of step 1. At the end of step 1, equation (6) implies that

$$\forall h = 1 \dots H \quad \sum_{k \in U_h / 0 < \pi_k^* < 1} \frac{\mathbf{x}_k}{\pi_k} \pi_k^* = \sum_{k \in U_h} \mathbf{x}_k - \sum_{k \in U_h / \pi_k^* = 1} \frac{\mathbf{x}_k}{\pi_k},$$

and summing these expressions yields

$$\sum_{k \in V} \frac{\mathbf{x}_k}{\pi_k} \pi_k^* = \sum_{k \in U} \mathbf{x}_k - \sum_{k \in U / \pi_k^* = 1} \frac{\mathbf{x}_k}{\pi_k}.$$

At the end of step 2, equation (6) leads to

$$\sum_{k \in V} \frac{\mathbf{x}_k}{\pi_k} \pi_k^{**} = \sum_{k \in V} \frac{\mathbf{x}_k}{\pi_k} \pi_k^*,$$

and combining the last two expressions, we get

$$\sum_{k \in V} \frac{\mathbf{x}_k}{\pi_k} \pi_k^{**} + \sum_{k \in U / \pi_k^* = 1} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k, \quad (9)$$

which ensures that balanced sampling on the variables \mathbf{x}_k is maintained exactly at the end of step 2. Step 3 completes the sampling process while maintaining the fixed-size constraint within each stratum U_h and can be carried out by means of a linear program to limit the lack of balance (see Deville and Tillé 2004).

The variance can be approximated with the variance formula proposed by Deville and Tillé (2005), if each flight phase in algorithm 1 is carried out with high entropy. Entropy can be increased substantially by performing a random sort on the population prior to sampling. In this case, the balancing variables are both the \mathbf{z}_k variables and the variables given by the product of the inclusion probabilities and the stratum membership indicators, which ensure a fixed sample size in each stratum. We have

$$\text{Var}(\hat{t}_{y\pi}) = \sum_{k \in U} \frac{b_k}{\pi_k^2} (y_k - \gamma' \mathbf{a}_k)^2 \quad (10)$$

with $\mathbf{a}_k = (\pi_k \mathbf{1}_{k \in U_1}, \dots, \pi_k \mathbf{1}_{k \in U_H}, \mathbf{z}_k')'$ and

$$\gamma = \left(\sum_{l \in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l \in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

We can use the variance estimator

$$v(\hat{t}_{y\pi}) = \sum_{k \in S} \frac{b_k}{\pi_k^3} (y_k - \hat{\gamma}' \mathbf{a}_k)^2 \quad (11)$$

proposed by Deville and Tillé (2005, page 578), with

$$\hat{\gamma} = \left(\sum_{l \in S} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l \in S} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

As shown in the variance approximation formula (10), it is important to note that the independence of the samples

from the various strata is lost with the proposed stratified balanced sampling method. The samples from strata U_1, \dots, U_H are coordinated to ensure overall balance across the whole population, which strips them of their independence. The Horvitz-Thompson estimator $\hat{t}_{y\pi}^h$ of total $t_{y\pi}$ remains unbiased. Its approximate variance is derived from equation (10) by replacing y_k with $y_k 1_{k \in U_h}$, and is given by

$$\text{Var}(\hat{t}_{y\pi}^h) \approx \sum_{k \in U} \frac{b_k}{\pi_k^2} (y_k 1_{k \in U_h} - (\gamma^h)' \mathbf{a}_k)^2 \quad (12)$$

with $\mathbf{a}_k = (\pi_k 1_{k \in U_1}, \dots, \pi_k 1_{k \in U_H}, \mathbf{z}_k')'$ and

$$\gamma^h = \left(\sum_{l \in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l \in U_h} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

In the particular case where the inference does not apply to the entire population but to a domain D that consists of a small number of strata, balanced sampling overall on the \mathbf{z} variables will be of little benefit. The variance of the Horvitz-Thompson estimator $\hat{t}_{y\pi}^D$ of total t_y^D for variable y for that domain will be close to the variance for stratified sampling, which is given by equation (1).

3. Quantitative results

In this section, we carry out a brief simulation study to test the performance of our sampling algorithm. First, we generate a finite population of 1,000, partitioned into 25 strata of equal size containing four variables: two variables of interest, y_1 and y_2 ; and two auxiliary variables, x_1 and x_2 . Variables x_1 and x_2 are generated with a gamma distribution with parameters 4 and 25. Variable y_1 is generated within stratum U_h using the model

$$y_1 = \alpha_{1h} + \varepsilon_h. \quad (13)$$

The ε_h are generated with a normal distribution with mean 0 and variance σ_h^2 . The model used to generate the values of y_1 is given by (13), with $\alpha_{1h} = 20h$ and variance σ_h^2 selected to produce a coefficient of determination R^2 approximately equal to 0.60 in each stratum. Variable y_2 is generated with the model

$$y_2 = \alpha_2 + \beta_2 x_1 + \gamma_2 x_2 + \eta. \quad (14)$$

The η are generated with a normal distribution with mean 0 and variance ρ^2 . The model used to generate the values of y_2 is given by (14), with $\alpha_2 = 500$, $\beta_2 = \gamma_2 = 5$, and variance ρ^2 selected to produce a coefficient of determination R^2 approximately equal to 0.60.

We are interested in estimating the total of variables y_1 and y_2 . We select a sample of $n = 25$ ($n = 50$ respectively) units with equal probabilities using three sampling designs:

Design 1: Stratified simple random sampling in each stratum

Design 2: Sampling balanced on variables π , x_1 and x_2

Design 3: Stratified sampling balanced on variables π , x_1 and x_2 , with pooling of the landing phases

In the case of stratified sampling, we have an allocation of size 1 (2 respectively) in each stratum. In the balanced designs, each flight phase is preceded by a random sort of the population. The variance associated with design 1 is calculated directly. The variance associated with designs 2 and 3 is approximated on the basis of 10,000 simulations. The results are presented in Table 1.

Table 1
Variance associated with the estimate of the total of two variables for a stratified design, a balanced design and a stratified balanced design with pooling of landing phases

Method	$n = 25$		$n = 50$	
	Total var.	Total var.	Total var.	Total var.
	y_1 ($\times 10^8$)	y_2 ($\times 10^9$)	y_1 ($\times 10^8$)	y_2 ($\times 10^9$)
Design 1	6.05	7.13	2.95	3.48
Design 2	14.31	3.05	7.02	1.40
Design 3	6.00	3.63	2.98	1.54

In each case, the proposed sampling design is comparable with the better of the two strategies. If the variable of interest is approximately constant across all strata, the proposed algorithm produces the same results as the stratified design. If the balancing variables are highly explanatory, the results produced by our algorithm and by direct balanced sampling are equivalent. The slight loss of precision comes from the landing phase: in the case of direct balanced sampling, we attempt to complete the sampling while limiting the lack of balance. With the proposed algorithm, the selected solution is suboptimal because we are imposing the additional constraint of a fixed size in each stratum.

In the case of stratified balanced sampling with pooling of the landing phases, Table 2 shows the variance given by 10,000 simulations and the variance given by the approximation formula (10).

Table 2
Comparison of the variance given by 10,000 simulations and the variance given by the approximation formula in the case of the estimation of two totals for a stratified balanced sampling design with pooling of landing phases

	$n = 25$		$n = 50$	
	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)
Simulation var.	6.0	3.6	3.0	1.5
Approximated var.	5.9	2.7	2.9	1.3

The approximation formula proposed by Deville and Tillé (2005) is close to exact if the variance associated with the landing phase is small relative to the variance associated with the flight phase. In the case of the y_2 variable, the balancing variables are highly explanatory. The variance is therefore larger for the landing phase than for the flight phase, and the approximation formula understates the actual variance. The variance associated with the landing phase will be considered in future studies.

Acknowledgements

The author is grateful to the referees and an associate editor for their constructive comments and suggestions.

References

- Bourdalle, G., Christine, M. and Wilms, L. (2000). Échantillons maître et emploi. Série INSEE Méthodes, Paris, France, 21, 139-173.
- Chauvet, G., and Tillé, Y. (2005). New SAS macros for balanced sampling. INSEE, Journées de Méthodologie Statistique, Paris.
- Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.
- Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.
- Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.
- Matei, A., and Tillé, Y. (2006). The R 'sampling' package. *European Conference on Quality in Survey Statistics*, Cardiff.
- Rousseau, S., and Tardieu, F. (2004). *La macro SAS CUBE d'échantillonnage équilibré - Documentation de l'utilisateur*. Technical report, INSEE, France.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 24, No. 4, 2008

Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities Joseph Sedransk	495
Discussion Nathaniel Schenker, Trivellore E. Raghunathan	507
Discussion David A. Binder	513
Model Averaging Methods for Weight Trimming Michael R. Elliott	517
A Note on the Asymptotic Equivalence of Jackknife and Linearization Variance Estimation for the Gini Coefficient Yves G. Berger	541
On Some Common Practices of Systematic Sampling Li-Chun Zhang	557
Does a Final Coverage Check Identify and Reduce Census Coverage Errors? Elizabeth Martin, Don A. Dillman	571
Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse Nancy Bates, James Dahlhamer, Eleanor Singer	591
Seasonality in Investment Plans and Their Revisions Alex Teterukovsky	612
Editorial Collaborators	623

Contents
Volume 25, No. 1, 2009

Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models Michael R. Elliott	1
The Use of Sample Weights in Hot Deck Imputation Rebecca R. Andridge, Roderick J. Little	21
Small Area Estimation in the Presence of Correlated Random Area Effects Monica Pratesi, Nicola Salvati	37
Nonparametric Variance Estimation for Nearest Neighbor Imputation Jun Shao	55
On the Inter-Regional Mover Problem in Panel Household Surveys Takis Merkouris	63
Compensating for Noncoverage of Nontelephone Households in Random-Digit-Dialing Surveys: A Comparison of Adjustments Based on Propensity Scores and Interruptions in Telephone Service K.P. Srinath, Martin R. Frankel, David C. Hoaglin, Michael P. Battaglia	77
Population Shifts and Demographic Methods C. Matthew Snipp, Juanita Tamayo Lott	99
Privacy-Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products Alan F. Karr, Xiaodong Lin, Ashish P. Sanil, Jerome P. Reiter	125
Using a Weighted Average of Base Period Price Indexes to Approximate a Superlative Index Janice Lent, Alan H. Dorfman	139
Book and Software Reviews Jana Asher, Trent D. Buskirk, John Hall, Gordon Willis	151
In Other Journals	165

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 36, No. 3, September/septembre 2008

Christopher A. FIELD Modeling biological data: several vignettes	341
Marco A.R. FERREIRA & Marc A. SUCHARD Bayesian analysis of elapsed times in continuous-time Markov chains	355
Amélie FILS-VILLETARD, Armelle GUILLOU & Johan SEGERS Projection estimators of Pickands dependence functions	369
Simon GUILLOTTE & François PERRON A Bayesian estimator for the dependence function of a bivariate extreme-value distribution	383
Jesús LÓPEZ-FIDALGO, Raul MARTÍN-MARTÍN & Douglas P. WIENS Marginally restricted sequential D-optimal designs	397
Kerrie P. NELSON & Don EDWARDS On population-based measures of agreement for binary classifications	411
Lei NIE, Haitao CHU & Valeriy R. KOROSTYSHEVSKIY Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits	427
Jean-François PLANTE Nonparametric adaptive likelihood weights	443
Yichuan ZHAO & Hongkun WANG Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data	463

Mylène BÉDARD & Jeffrey S. ROSENTHAL Optimal scaling of Metropolis algorithms: heading toward general target distributions	483
Jörn DANNEMANN & Hajo HOLZMANN Testing for two states in a hidden Markov model.....	505
Christopher A. FIELD, Zhen PANG & Alan H. WELSH Bootstrapping data with multiple levels of variation	521
Feifang HU, Li-Xin ZHANG, Siu H. CHEUNG & Wai S. CHAN Doubly adaptive biased coin designs with delayed responses	541
Sai Man Simon KWOK & Wai Keung LI On diagnostic checking of the autoregressive conditional intensity model	561
Omer OZTURK Inference in the presence of ranking error in ranked set sampling	577
Matías SALIBIÁN-BARRERA & Ying WEI Weighted quantile regression with nonelliptically structured covariates.....	595
Ana-Maria STAICU & Nancy M. REID On probability matching priors.....	613
Robin WILLINK Shrinkage confidence intervals for the normal mean: using a guess for greater efficiency.....	623

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférentiellement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1.	Présentation	1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 ½ pouce tout autour. 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte. 1.4 Les remerciements doivent paraître à la fin du texte. 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.	Résumé	Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3.	Rédaction	3.1 Éviter les notes au bas des pages, les abréviations et les sigles. 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(.) et log(.) etc. 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin. 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique. 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, I). 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4.	Figures et tableaux	4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
5.	Bibliographie	5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164). 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6.	Communications brèves	6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

My��ne B��DARD & Jeffrey S. ROSENTHAL	Optimal scaling of Metropolis algorithms: heading toward general target distributions.....	483
J��m DANNEMANN & Hajo HOLZMANN	Testing for two states in a hidden Markov model.....	505
Christopher A. FIELD, Zhen PANG & Alan H. WELSH	Bootstrapping data with multiple levels of variation.....	521
Feifang HU, Li-Xin ZHANG, Siu H. CHEUNG & Wai S. CHAN	Doubly adaptive biased coin designs with delayed responses.....	541
Sai Man Simon KWOK & Wai Keung LI	On diagnostic checking of the autoregressive conditional intensity model.....	561
Omer OZTURK	Inference in the presence of ranking error in ranked set sampling.....	577
Matias SALIBA��N-BARRERA & Ying WEI	Weighted quantile regression with nonelliptically structured covariates.....	595
Ana-Maria STACU & Nancy M. REID	On probability matching priors.....	613
Robin WILLINK	Shrinkage confidence intervals for the normal mean: using a guess for greater efficiency.....	623

Volume 36, No. 3, September/septembre 2008

Christoph A. FIELD	Modeling biological data: several vignettes	341
Marco A.R. FERREIRA & Marc A. SUCHARD	Bayesian analysis of elapsed times in continuous-time Markov chains	355
Amélie FILS-VILLETARD, Armelle GUILLOU & Johan SEGERS	Projection estimators of Pickands dependence functions	369
Simon GUILLOTTE & François PERON	A Bayesian estimator for the dependence function of a bivariate extreme-value distribution	383
Jesús LÓPEZ-FIDALGO, Raul MARTÍN-MARTÍN & Douglas P. WIENS	Marginally restricted sequential D-optimal designs	397
Kerrie P. NELSON & Don EDWARDS	On population-based measures of agreement for binary classifications	411
Lei NIE, Haitao CHU & Valery R. KOROSTYSHEVSKIY	Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits	427
Jean-François PLANTE	Nonparametric adaptive likelihood weights	443
Yichuan ZHAO & Hongkun WANG	Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data	463

Contents Volume 25, No. 1, 2009

Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models	Michael R. Elliott	1
The Use of Sample Weights in Hot Deck Imputation	Rebecca R. Andridge, Roderick J. Little	21
Small Area Estimation in the Presence of Correlated Random Area Effects	Monica Pratesi, Nicola Salvati	37
Nonparametric Variance Estimation for Nearest Neighbor Imputation	Jun Shao	55
On the Inter-Regional Mover Problem in Panel Household Surveys	Takis Merkouris	63
Compensating for Noncoverage of Nonteleshone Households in Random-Digit-Dialing Surveys: A Comparison of Adjustments Based on Propensity Scores and Interruptions in Telephone Service	K. P. Srinath, Martin R. Frankel, David C. Hoaglin, Michael P. Battaglia	77
Population Shifts and Demographic Methods	C. Matthew Snipp, Juanita Tamayo Lott	99
Privacy-Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products	Alan F. Karr, Xiaodong Lin, Ashish P. Sanil, Jerome P. Reiter	125
Using a Weighted Average of Base Period Price Indexes to Approximate a Superlative Index	Janice Lent, Alan H. Dorfman	139
Book and Software Reviews	Jana Asher, Trent D. Buskirk, John Hall, Gordon Willis	151
In Other Journals		165

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 24, No. 4, 2008

Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities	Joseph Sedransk	495
Discussion	Nathaniel Schenker, Trivellore E. Raghunathan	507
Discussion	David A. Binder	513
Model Averaging Methods for Weight Trimming	Michael R. Elliott	517
A Note on the Asymptotic Equivalence of Jackknife and Linearization Variance Estimation for the Gini Coefficient	Yves G. Berger	541
On Some Common Practices of Systematic Sampling	Li-Chun Zhang	557
Does a Final Coverage Check Identify and Reduce Census Coverage Errors?	Elizabeth Martin, Don A. Dillman	571
Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse	Nancy Bates, James Dahlander, Eleanor Singer	591
Seasonality in Investment Plans and Their Revisions	Alex Teterukovsky	612
Editorial Collaborators		623

Tableau 1
Variance associée à l'estimation du total de 2 variables pour les plans de sondage stratifié, équilibré, et stratifié équilibré avec mise en commun des phases d'atterrissage

Méthode	$n = 25$		$n = 50$	
	Var. total	Var. total	Var. total	Var. total
y_1 ($\times 10^8$)	6,05	7,13	2,95	3,48
y_2 ($\times 10^9$)	3,05	7,02	1,40	1,54
Plan 1	14,31			
Plan 2	6,00			
Plan 3	3,63			

Dans chaque cas, le plan de sondage proposé est comparable avec la meilleure des deux stratégies. Si la variable d'intérêt est approximativement constante par strate, l'algorithme proposé donne les mêmes résultats que le plan de sondage stratifié. Si les variables d'équilibrage sont bien explicatives, les résultats obtenus avec notre algorithme et avec un tirage équilibré direct sont équivalents. La légère perte de précision provient de l'étape d'atterrissage : dans le cas du tirage équilibré direct, on cherche à terminer l'échantillonnage en limitant le défaut d'équilibrage. Avec l'algorithme proposé, la solution retenue est sous-optimale car on ajoute la contrainte supplémentaire d'une taille fixe dans chaque strate.

Dans le cas du tirage stratifié équilibré avec mise en commun des phases d'atterrissage, le tableau 2 compare la variance donnée par 10 000 simulations avec la formule de variance approchée donnée en (10).

Tableau 2
Comparaison entre la variance donnée par 10 000 simulations et la formule de variance approchée dans le cas de l'estimation de deux totaux pour un plan de sondage stratifié équilibré avec mise en commun des phases d'atterrissage

$n = 25$		$n = 50$	
Total y_1 ($\times 10^8$)	6,0	Total y_1 ($\times 10^8$)	3,0
Total y_2 ($\times 10^9$)	3,6	Total y_2 ($\times 10^9$)	2,9
Var. simulations	5,9	Var. approchée	1,5
			1,3

Bibliographie

Remerciements
L'auteur remercie les examinateurs et un éditeur associé pour leurs commentaires et suggestions constructifs.

Bourdalle, G., Christine, M. et Wilms, L. (2000). Échantillons maître et emploi. Sène INSEE Méthodes, Paris, France, 21, 139-173.

Chauvet, G., et Tillé, Y. (2005). New SAS macros for balanced sampling. INSEE, Journées de Méthodologie Statistique, Paris.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Matei, A., et Tillé, Y. (2006). The R 'sampling' package. *European Conference on Quality in Survey Statistics*, Cardiff.

Roussseau, S., et Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré - Documentation de l'utilisateur. Rapport technique, INSEE, France.

de terminer l'échantillonnage en respectant la contrainte de taille fixe à l'intérieur de chaque strate U_h , et peut être réalisé au moyen d'un programme linéaire afin de limiter le défaut d'équilibrage (cf: Deville et Tillé 2004). La variance peut être approchée à l'aide de la formule de Deville et Tillé (2005), si chaque phase de vol de l'algorithme est réalisée avec une entropie forte. L'entropie peut être fortement augmentée en tirant aléatoirement la population concernée préalablement au tirage. Les variables d'équilibrage sont ici d'une part les variables \mathbf{z}_k , et d'autre part les variables données par le produit des probabilités d'inclusion et des indicatrices d'appartenance aux strates et assurant une taille fixe d'échantillon dans chaque strate. On a

$$(10) \quad \text{Var}(f_{y^n}) = \sum_{k \in U} \frac{\pi_k^2}{b_k} (Y_k - \gamma' \mathbf{a}_k)^2$$

avec $\mathbf{a}_k = (\pi_k 1_{k \in U_1}, \dots, \pi_k 1_{k \in U_H}, \mathbf{z}_k')$ et

$$\gamma = \left(\sum_{l \in U} b_l \frac{\mathbf{a}_l \mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l \in U} b_l \frac{\pi_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l}.$$

On peut utiliser l'estimateur de variance

$$(11) \quad v(f_{y^n}) = \sum_{k \in S} \frac{\pi_k^2}{b_k} (Y_k - \hat{\gamma}' \mathbf{a}_k)^2$$

proposé par Deville et Tillé (2005), page 578, avec

$$\hat{\gamma} = \left(\sum_{l \in S} b_l \frac{\pi_l}{\pi_l} \frac{\mathbf{a}_l \mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l \in S} b_l \frac{\pi_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l}.$$

Ainsi qu'il apparaît dans la formule (10) de variance approchée, il est important de noter que l'indépendance des tirages dans les différentes strates est perdue avec la méthode d'échantillonnage équilibrée stratifiée proposée. Les tirages dans les strates U_1, \dots, U_H sont en effet coordonnés de façon à assurer un équilibrage global sur l'ensemble de la population, ce qui supprime l'indépendance. L'estimateur de Horvitz-Thompson $f_{y^n}^H$ du total t_{y^n} reste sans biais. Sa variance approchée se déduit de la formule (10) en remplaçant Y_k par $\gamma_k 1_{k \in U_h}$ et est donnée par

$$(12) \quad \text{Var}(f_{y^n}^H) \approx \sum_{k \in U} \frac{\pi_k^2}{b_k} (Y_k 1_{k \in U_h} - (\gamma_h)' \mathbf{a}_k)^2$$

avec $\mathbf{a}_k = (\pi_k 1_{k \in U_1}, \dots, \pi_k 1_{k \in U_H}, \mathbf{z}_k')$ et

$$\gamma_h = \left(\sum_{l \in U_h} b_l \frac{\pi_l}{\pi_l} \frac{\mathbf{a}_l \mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l \in U_h} b_l \frac{\pi_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l}.$$

Dans le cas particulier où l'inférence ne porte pas sur la population entière mais sur un domaine D comportant un nombre limité de strates, le bénéfice de l'équilibrage global

sur les variables \mathbf{z} sera faible. La variance de l'estimateur de Horvitz-Thompson $f_{y^n}^H$ du total t_{y^n} de la variable y sur ce domaine sera proche de la variance obtenue dans le cas d'un échantillonnage stratifié et donnée par la formule (1).

3. Résultats numériques

Nous réalisons une courte étude par simulations pour tester les performances de notre algorithme d'échantillonnage. Nous générons tout d'abord une population finie de taille 1 000, partitionnée en 25 strates de même taille, et contenant 4 variables : 2 variables d'intérêt y_1 et y_2 , et 2 variables auxiliaires x_1 et x_2 . Tout d'abord, les variables x_1 et x_2 sont générées selon une distribution Gamma de paramètres 4 et 25. La variable y_1 est générée au sein de la strate U_h selon le modèle

$$(13) \quad y_1 = \alpha_{1h} + \varepsilon_{1h}.$$

Les ε_{1h} sont générés selon une distribution normale de moyenne 0 et de variance σ_1^2 . Le modèle utilisé pour générer les valeurs de y_1 est donné par (13), avec $\alpha_{1h} = 20h$ et une variance σ_1^2 choisie pour donner un coefficient de détermination R^2 approximativement égal à 0,60 au sein de chaque strate. La variable y_2 est générée selon le modèle

$$(14) \quad y_2 = \alpha_2 + \beta_2 x_1 + \gamma_2 x_2 + \eta.$$

Les η sont générés selon une distribution normale de moyenne 0 et de variance σ_2^2 . Le modèle utilisé pour générer les valeurs de y_2 est donné par (14), avec $\alpha_2 = 500$, $\beta_2 = \gamma_2 = 5$, et une variance σ_2^2 choisie pour donner un coefficient de détermination R^2 approximativement égal à 0,60.

On s'intéresse à l'estimation du total des variables y_1 et y_2 . On sélectionne un échantillon de $n = 25$ (respectivement $n = 50$) unités à probabilités égales selon trois plans de sondage :

Plan 1 : tirage stratifié avec sondage aléatoire simple dans chaque strate,
Plan 2 : tirage équilibré sur les variables π, x_1 et x_2 ,
Plan 3 : tirage équilibré sur les variables π, x_1 et x_2 , stratifié avec mise en commun des phases d'atterrissage.

Dans le cas du tirage stratifié, on a donc une allocation de taille 1 (respectivement 2) dans chaque strate. Dans les tirages équilibrés, chaque phase de vol est précédée d'un tirage aléatoire de la population concernée. La variance associée au plan 1 est calculée directement. La variance associée aux plans 2 et 3 est approchée sur la base de 10 000 simulations. Le tableau 1 compare les résultats obtenus.

d'échantillon dans chaque strate. Dans certains cas, cette contrainte ne peut être respectée. Il est en effet fréquent d'utiliser un découpage très fin de la population afin d'améliorer sa pertinence, ce qui revient à réduire la taille d'échantillon sélectionnée dans chaque strate en se fixant généralement la limite d'un échantillon de taille 2 pour chacune afin de pouvoir obtenir un estimateur sans biais de variance.

Nous nous plaçons à nouveau dans le cas d'une population U découpée en H strates U_1, \dots, U_H , et dans laquelle un vecteur $\mathbf{x}_k = (\pi_k^1 \mathbf{z}_k^1, \dots, \pi_k^H \mathbf{z}_k^H)'$ de variables auxiliaires est connu. On suppose que la variable π_k fait partie des contraintes d'équilibrage, afin d'assurer un échantillonnage de taille fixe. Dans le cas où la répartition par strate est trop faible pour que celle de taille fixe dans chaque strate, l'algorithme 1 fournit une méthode alternative d'échantillonnage. Une phase de vol est réalisée indépendamment sur chacune des H strates : on note $\pi_k^h = (\pi_k^h)^{k \in U_h}$, $h = 1, \dots, H$ les vecteurs de probabilités obtenus à l'issue de ces phases de vol, $\pi^* = (\pi_k^1)^{k \in U_1}, \dots, \pi_k^H$ désigne les unités qui n'ont pas encore été échantillonnées ou rejetées, et $\mathbf{x}_k^* = (\pi_k^1 \mathbf{z}_k^1)^{k \in U_1}, \dots, \pi_k^H \mathbf{z}_k^H$. Le vecteur des probabilités obtenues après une dernière phase de vol sur l'ensemble de ces unités restantes est noté $\pi^{**} = (\pi_k^{**})^{k \in U}$. L'ensemble des unités de la strate U_h qui n'ont pas encore été échantillonnées ou rejetées à l'issue de cette nouvelle phase de vol est noté W_h .

Algorithme 1 : Échantillonnage équilibré stratifié avec mise en commun des phases d'ajustements

Etape 1. Réaliser une phase de vol, avec les variables d'équilibrage \mathbf{x}_k et les probabilités d'inclusion π_k , indépendamment dans chaque strate U_h .

Etape 2. Réaliser une phase de vol, avec les variables d'équilibrage \mathbf{x}_k^* et les probabilités d'inclusion π_k^* , sur l'ensemble V des unités restantes à l'issue de l'étape 1.

Etape 3. Sélectionner un échantillon de taille fixe dans chaque sous-population W_h , avec des probabilités d'inclusion π_k^{**} .

L'algorithme s'inspire d'une méthode utilisée à l'Institut National de la Statistique et des Études Économiques (Insee) pour la sélection des unités primaires de l'échantillon-maître 1999. L'échantillon-maître est un échantillon de logements, sélectionné dans le Recensement de 1999, et servant de base de sondage pour les enquêtes sur les ménages. On trouvera une description détaillée du plan de sondage de l'échantillon-Maître dans Bourdelle, Christine et Wilms (2000). Les logements sont à l'origine regroupés au sein d'unités urbaines ou d'unités rurales. Dans la sous-population des unités de moins de 100 000

habitants, un échantillon d'environ 6 % est sélectionné. On dispose de quatre variables auxiliaires (revenu net imposable, et effectif selon trois tranches d'âge). Le nombre attendu d'unités échantillonnées est trop faible pour permettre un tirage stratifié selon la région, avec équilibrage selon ces quatre variables dans chaque région. Les régions ont donc été regroupées en 8 super-régions, et les tirages sont coordonnés de manière à assurer, d'une part, un équilibrage global pour les quatre variables auxiliaires sur chaque super-région, et d'autre part une taille fixe d'échantillon dans chaque région.

Une méthode similaire est également proposée par Rousseau et Tardieu (2004) pour la sélection d'échantillons équilibrés dans de grandes bases de sondage en utilisant la macro CUBE disponible sur le site Internet de l'Insee. Le temps d'exécution de cette macro est en effet approximativement proportionnel au carré de la taille de la population. Notons que Chauvet et Tillé (2006) proposent une méthode rapide d'échantillonnage équilibré dont le temps de calcul ne dépend plus que de la taille de la population, et permet de sélectionner directement des échantillons équilibrés sur de très grandes populations. L'algorithme a été programmé sous forme d'une macro SAS, voir Chauvet et Tillé (2005), et il est également disponible dans le R Sampling Package de Malet et Tillé (2006). Dans chacun des deux programmes, la seconde phase de vol est réalisée en ajoutant aux variables d'équilibrage \mathbf{x}_k^* une contrainte associée à chaque strate, et permettant de maintenir la condition de taille fixe dans chacune d'entre elles.

Utiliser le vecteur de probabilités d'inclusion π^* conditionnellement au résultat obtenu à l'issue de l'étape 1, assure que le vecteur π des probabilités d'inclusion est respecté en déconditionnant par rapport au résultat de l'étape 1. À l'issue de cette 1^{ère} étape, l'équation (6) implique que

$$VH = 1 \dots H \quad \sum_{k \in U_h / 0 < \pi_k^* < 1} \frac{\pi_k}{\mathbf{x}_k} \pi_k^* = \sum_{k \in U_h} \frac{\pi_k}{\mathbf{x}_k} - \sum_{k \in U_h / \pi_k^* = 1} \frac{\pi_k}{\mathbf{x}_k}$$

et en sommant ces expressions

$$\sum_{k \in V} \frac{\pi_k}{\mathbf{x}_k} \pi_k^* = \sum_{k \in U} \frac{\pi_k}{\mathbf{x}_k} - \sum_{k \in U / \pi_k^* = 1} \frac{\pi_k}{\mathbf{x}_k}.$$

À l'issue de l'étape 2, on obtient à l'aide de l'équation (6)

$$\sum_{k \in V} \frac{\pi_k}{\mathbf{x}_k} \pi_k^{**} = \sum_{k \in V} \frac{\pi_k}{\mathbf{x}_k} \pi_k^*.$$

et en comparant ces deux dernières expressions

$$(9) \quad \sum_{k \in V} \frac{\pi_k}{\mathbf{x}_k} \pi_k^{**} + \sum_{k \in U / \pi_k^* = 1} \frac{\pi_k}{\mathbf{x}_k} = \sum_{k \in U} \frac{\pi_k}{\mathbf{x}_k}$$

ce qui assure que l'équilibrage sur les variables \mathbf{x}_k est exactement respecté à l'issue de l'étape 2. L'étape 3 permet

$$(4) \quad f_{h^*}^{\pi} = f_{h^*}^{\pi^*}$$

La méthode du Cube (Deville et Tillé (2004) permet de sélectionner des échantillons approximativement équilibrés sur un nombre quelconque de variables, en respectant exactement un jeu de probabilités d'inclusion π préalablement choisi. Elle se décompose en deux phases appelées phase de vol et phase d'atterrissage. À chaque étape de la phase de vol, on décide aléatoirement de sélectionner ou d'écarter définitivement l'une des unités de la population. À l'issue de la phase de vol, on obtient dans chaque strate U_h un vecteur $\pi^{h*} = (\pi_k^{h*})_{k \in U_h}$, $\pi_k^{h*} \in [0, 1]^N$ vérifiant les conditions suivantes :

$$(5) \quad E(\pi^{h*}) = \pi^h,$$

$$(6) \quad \sum_{k \in U_h} \frac{\pi_k}{\pi_k^*} \pi_k^* = \sum_{k \in U_h} \pi_k^*,$$

$$(7) \quad \text{Card}\{k \in U_h; 0 < \pi_k^* < 1\} \leq q,$$

où E désigne l'espérance sous le mécanisme d'échantillonnage associé à la phase de vol. Le vecteur π^{h*} donne le résultat de la phase de vol : π_k^* vaut 1 si l'unité k est sélectionnée, 0 si elle est rejetée et est comprise entre 0 et 1 strictement si la décision n'est pas encore prise pour l'unité k après la phase de vol. Les équations (5) et (6) assurent que les probabilités d'inclusion et les contraintes d'équilibrage sont parfaitement respectées à l'issue de la phase de vol. L'équation (7) assure qu'il reste à trancher pour au plus q individus dans chaque strate U_h , où q désigne le nombre de variables d'équilibrage. La phase de vol s'arrête quand les contraintes d'équilibrage ne peuvent plus être exactement respectées. La phase d'atterrissage consiste alors à définir, conditionnellement au résultat de la phase de vol, un plan de sondage optimal défini sur la population N des unités restantes. Ce plan est optimal au sens où il permet de terminer l'échantillonnage en minimisant la variance conditionnelle à la phase de vol de l'estimateur de Horvitz-Thompson des variables d'équilibrage. Les unités restantes sont échantillonnées, conditionnellement au résultat de la phase de vol, avec les probabilités d'inclusion $(\pi_k^{k_{ev}})$, de sorte que les probabilités d'inclusion non conditionnelles $(\pi_k^{k_{ev}})$ de ces unités soient exactement respectées.

La mesure d'entropie associée à un plan de sondage (p) défini sur la population N est donnée par

$$I(p) = - \sum_{s \subset N} p(s) \log(p(s)),$$

avec la convention $0 \log(0) = 0$. Deville et Tillé (2005) montrent que le plan équilibré à entropie maximale parmi les plans de sondage équilibrés sur les mêmes variables et respectant les mêmes probabilités d'inclusion peut être vu

comme le conditionnel d'un plan de Poisson. Sous une hypothèse de normalité asymptotique de l'estimateur de Horvitz-Thompson multivarié dans le cas d'un plan poissonien, ils en déduisent une formule approchée de variance pour l'estimateur de Horvitz-Thompson pour un plan de sondage équilibré. Dans le cas de l'échantillonnage équilibré stratifié, on obtient

$$(8) \quad \text{Var}(f_{y^{\pi}}) = \sum_{h=1}^H \sum_{k \in U_h} \frac{\pi_k^2}{b_k^2} (y_k - \beta_h \mathbf{x}_k)^2$$

où $\beta_h = (\sum_{k \in U_h} b_k \mathbf{x}_k / \pi_k)^{-1} \sum_{k \in U_h} b_k \mathbf{x}_k / \pi_k y_k / \pi_k$. Deville et Tillé (2005) proposent plusieurs approximations pour les coefficients b_k . La plus simple consiste à utiliser $b_k = \pi_k(1 - \pi_k)$. La variance de l'estimateur de Horvitz-Thompson sera faible si dans chaque strate la variable d'intérêt y est bien expliquée par les variables d'équilibrage \mathbf{x} .

L'équilibrage sera bien respecté au sein de chaque strate si le nombre de variables d'équilibrage reste faible devant la taille d'échantillon. Mais dans certains cas, la répartition par strate est trop faible pour permettre l'équilibrage : si la population est stratifiée de façon très fine, une pratique courante consiste à sélectionner un échantillon de taille 2 dans chaque strate. Il n'est alors pas possible d'imposer une condition autre que la contrainte de taille fixe d'échantillon dans chaque strate.

Nous proposons dans la section suivante un algorithme d'échantillonnage adapté de la méthode du Cube, assurant un équilibrage sur l'ensemble de la population pour des variables choisies et permettant le respect strict de la répartition souhaitée au sein de chaque strate. Les échantillons ne sont alors plus sélectionnés indépendamment dans chaque strate. La précision est améliorée par rapport à un sondage stratifié avec tirage de taille fixe dans chaque strate si sur l'ensemble de la population les variables d'équilibrage sont bien corrélées à la variable d'intérêt. Cet algorithme présente également l'avantage de permettre un équilibrage approximatif dans chaque strate, qui sera d'autant mieux respecté que la taille d'échantillon qui y est allouée est importante.

2. Échantillonnage équilibré stratifié avec mise en commun des phases d'atterrissage

Si l'échantillon S est sélectionné dans N selon la procédure d'échantillonnage équilibré stratifié présentée en section 1, l'équilibrage sera bien respecté au sein de chaque strate si la phase d'atterrissage porte sur un faible nombre d'individus par rapport à la taille d'échantillon. Plus précisément, l'équation (7) montre que le nombre de variables d'équilibrage doit être faible par rapport à la répartition

Echantillonnage équilibré stratifié

Guillaume Chauvet¹

Résumé

Lors de la sélection d'un échantillon, une pratique courante consiste à définir un plan de sondage stratifié sur des sous-populations. La variance de l'estimateur de Horvitz-Thompson est alors réduite par rapport à un tirage direct si les strates sont bien homogènes au regard de la variable d'intérêt. Si des variables auxiliaires sont disponibles pour chaque individu, l'échantillonnage peut être amélioré par tirage équilibré au sein de chaque strate et l'estimateur de Horvitz-Thompson sera plus précis si les variables auxiliaires sont bien corrélées à la variable d'intérêt. Cependant, si la répartition d'échantillon est faible dans certaines strates, l'équilibrage ne sera respecté que de façon très approximative. Nous proposons ici une méthode de tirage permettant de sélectionner un échantillon équilibré sur l'ensemble de la population, en respectant une allocation fixée au sein de chaque strate. Nous montrons que dans le cas particulier important d'un tirage de taille 2 dans chaque strate, la précision de l'estimateur de Horvitz-Thompson est améliorée si la variable d'intérêt est bien expliquée par les variables d'équilibrage sur l'ensemble de la population. Une application au cas d'un échantillonnage rotatif est également proposée.

Mots clés : Échantillonnage rotatif ; entropie maximale ; méthode du Cube ; stratification ; tirage à probabilités intégrales.

1. Introduction

Dans le cas d'un tirage stratifié, la population U est partitionnée en H sous-populations U_h , $h = 1, \dots, H$ appelées strates, dans lesquelles des échantillons S_h , $h = 1, \dots, H$ sont respectivement sélectionnés selon des plans de sondage indépendants p_h , $h = 1, \dots, H$. La

probabilité d'inclusion de l'unité k est la probabilité π_k que l'unité k soit sélectionnée dans l'échantillon, et la probabilité d'inclusion jointe est la probabilité $\pi_{k\ell}$ que deux unités distinctes k et ℓ soient sélectionnées conjointement dans l'échantillon. On notera $\pi = (\pi_k)_{k \in U}$ et $\pi'' = (\pi_{k\ell})_{k, \ell \in U}$. On suppose qu'au sein de chaque strate U_h , le plan $p_h(\cdot)$ est de taille fixe. On a alors en particulier $\sum_{k \in U_h} \pi_k = n_h$, $h = 1, \dots, H$ où n_h désigne la répartition dans la strate U_h . Nous supposons dans la suite de l'article que

toutes les tailles d'échantillon par strate n_h sont entières. L'estimateur de Horvitz-Thompson $\hat{t}_{zt} = \sum_{k \in S} \mathbf{z}_k / \pi_k = \sum_{h=1}^H \hat{t}_{zt}^h$ ou $\hat{t}_{zt}^h = \sum_{k \in S_h} \mathbf{z}_k / \pi_k$, estime sans biais $t_z = \sum_{k \in U} \mathbf{z}_k$ ou $t_z^h = \sum_{k \in U_h} \mathbf{z}_k$ désigne le total sur U_h de la variable (vectorielle) \mathbf{z} . Dans le cas particulier où $\mathbf{z}_k = y_k$ est scalaire, la variance de l'estimateur de Horvitz-Thompson s'obtient à l'aide de la formule de variance de

Cette variance sera faible si les strates sont homogènes au regard de la variable d'intérêt, plus exactement si y_k / π_k est approximativement constant au sein de chaque strate.

$$\text{Var}(\hat{t}_{yt}) = \frac{1}{H} \sum_{h=1}^H \sum_{k \neq \ell \in U_h} \left(\pi_k \pi_\ell - \pi_{k\ell} \right) \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2. \quad (1)$$

Sen-Yates-Grundy :

Le tirage stratifié de taille fixe dans chaque strate est donc un cas particulier d'échantillonnage équilibré. Dans le cas d'un nombre quelconque de contraintes, un échantillon exactement équilibré ne peut généralement être trouvé. Supposons par exemple que la population U_h contienne 100 individus sur lesquels est définie une variable x à deux modalités, 0 et 1, telle que 53 individus de la population présentent la modalité 0. Sélectionner un échantillon de taille 10, à probabilités égales, équilibré sur la variable x , supposerait de sélectionner un échantillon contenant 5,3 individus présentant la modalité $x = 0$ et 4,7 individus présentant la modalité $x = 1$, ce qui est impossible. L'objectif est donc généralement de sélectionner un échantillon approximativement équilibré, c'est-à-dire tel que

$$\sum_{k \in S_h} 1 = \sum_{k \in U_h} \pi_k = n_h. \quad (3)$$

variables \mathbf{x} si les équations $t_{\mathbf{x}}^h = t_{\mathbf{x}}^h$ sont exactement respectées. La variance de l'estimateur de Horvitz-Thompson est donc nulle pour l'estimation du total des variables d'équilibrage. Dans le cas particulier où $\mathbf{x} = \pi$, c'est-à-dire si la probabilité d'inclusion est la seule variable d'équilibrage, les équations (2) se réduisent à

Si un vecteur $\mathbf{x} = (x_1, \dots, x_q)$ de q variables auxiliaires est disponible avant tirage pour chaque individu de la population, l'échantillonnage au sein de chaque strate peut être amélioré à l'aide de l'algorithme du Cube (Dewille et Tillé 2004) qui permet de sélectionner des échantillons équilibrés. Le plan de sondage $p_h(\cdot)$ est dit équilibré sur les

- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Sämdal, C., et Lundström, S. (2005). Estimation in Surveys with Nonresponse. *Wiley Series in Survey Methodology*, John Wiley & Sons, Chichester, England.
- Sämdal, C., Swensson, B. et Wretman, J. (2003). Model-assisted survey sampling. *Springer Series in Statistics*, Springer, New York.
- Schouen, B. (2004). Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998. *Papier de discussion cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2004/default.htm*.

- Schouen, B., et Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes. *Article de discussion 07002*, CBS, Voorburg. Disponible au <http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2007/default.htm>.
- Stoop, I. (2005). Surveying nonrespondents. *Field Methods*, 16, 23-34.

- Voogt, R. (2004). I am not interested: Nonresponse bias, response bias and stimulus effects in election research. *Thèse de doctorat*, University of Amsterdam, Amsterdam.

prédisent le biais de non-réponse des variables étudiées dépend du mécanisme de création de données manquantes. Si celui-ci est fortement non ignorable, les indicateurs R ne donneront pas de bons résultats. Cependant, sans aucun renseignement sur le mécanisme de données manquantes, aucun autre indicateur ne le ferait non plus. C'est pourquoi nous avons établi la notion de biais absolu maximal, car elle donne une limite au biais de non-réponse sous le pire scénario. Un deuxième sujet de futurs travaux de recherche est le calcul théorique de l'erreur-type de l'indicateur R utilisé dans le présent article. Les erreurs obtenues par le bootstrap non paramétrique ne donnent que des approximations naïves. Or, si nous voulons que les indicateurs R jouent un rôle plus actif dans la comparaison de diverses stratégies, nous devons obtenir des formes (approximativement) explicites. Troisièrement, nous devons étudier la relation entre le choix et le nombre de variables auxiliaires, d'une part, et les erreurs-types de l'indicateur R, d'autre part.

Remerciements

Les auteurs remercient Bob Groves, Björn Janssen, Geert Loosveldt, ainsi que le rédacteur associé et deux examinateurs de leurs suggestions et commentaires constructifs.

Bibliographie

Agresti, A. (2002). Categorical data analysis. *Wiley Series in Probability and Statistics*. New York : John Wiley & Sons, Inc., NY, USA.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 2, 238-246.

Berthoin, S. (2006). A measure of representativeness of a sample for inferential purposes. *Revue Internationale de Statistique*, 74, 149-159.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 3, 251-260.

Cobben, F., Janssen, B., Berkel, K. van et Brakel, J. van den (2007). Statistical inference in a mixed-mode data collection setting. Document présenté au ISI 2007, 23 au 29 août 2007, Lisbon, Portugal.

Cobben, F., et Schouten, B. (2007). An empirical validation of R-indicators. Papier de discussion, CBS, Voorburg.

Curtin, R., Presser, S. et Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-428.

Efron, B., et Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Fouwels, S., Janssen, B. et Weitzels, W. (2006). Experiment mixed-mode waarneming bij de VMR. Rapport technique SOO-2007-H53, CBS, Heerlen.

Goodman, L.A., et Kruskal, W.H. (1979). Measures of association for cross-classifications. Springer-Verlag, Berlin, Duitsland.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 5, 646-675.

Groves, R.M., Presser, S. et Dippo, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68, 2-31.

Häjek, J. (1981). Sampling from finite populations. New York : Marcel Dekker, USA.

Hansen, M.H., et Hurwitz, W.H. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Heerwegh, D., Abis, K. et Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 1, 3-10.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 2, 119-127.

Keefer, S., Miller, C., Kohut, A., Groves, R.M., et Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.

Kerstem, H.M.P., et Bethlehem, J.G. (1984). Exploring an reducing the nonresponse bias by asking the basic question. *Statistical Journal of the United Nations*, ECE 2, 369-380.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.

Kruskal, W., et Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *Revue Internationale de Statistique*, 47, 13-24.

Kruskal, W., et Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *Revue Internationale de Statistique*, 47, 111-123.

Kruskal, W., et Mosteller, F. (1979c). Representative sampling III: Current statistical literature. *Revue Internationale de Statistique*, 47, 245-265.

Little, R.J.A., et Rubin, D.B. (2002). Statistical analysis with missing data. Wiley Series in Probability and Statistics. New York : John Wiley & Sons, Inc., NY, USA.

Marsh, H.W., Balla, J.R. et McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 3, 391-410.

Merkle, D.M., et Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. Dans *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little). New York : John Wiley & Sons, Inc., 243-258.

Les résultats présentés au tableau 6 ne contredisent pas ceux de Fowels et coll. (2006). Nous constatons aussi que la composition de la réponse en ligne à l'essai pilote est moins équilibrée, tandis que celle de la réponse complète à l'essai pilote n'est pas sensiblement moins bonne que celle de la réponse à l'enquête de surveillance de la sécurité proprement dite.

6. Discussion

Nous poursuivons trois grands objectifs dans le présent

article : donner une définition mathématiquement rigoureuse et une perception de la représentativité de la réponse, construire un indicateur possible de la représentativité et illustrer empiriquement l'utilisation de l'indicateur. Comme nous l'avons vu, l'indicateur proposé est un exemple de ce que nous appelons les indicateurs R , où « R » signifie représentativité. Au moyen de l'exemple empirique, nous cherchons à appuyer l'idée que ces indicateurs R sont des outils valables pour comparer différentes enquêtes ou stratégies de collecte des données. Les indicateurs R sont utiles s'ils permettent de confirmer les résultats d'analyses complexes effectuées dans le cadre d'études ayant trait à des enquêtes réalisées sur de multiples périodes ou à de multiples enquêtes portant sur un sujet particulier.

L'indicateur R décrit dans le présent article est prometteur, parce qu'il est facile à calculer et peut être interprété et normalisé quand les propositions à répondre peuvent être estimées sans erreur. L'application à des données d'enquête réelles montre que l'indicateur R confirme les analyses antérieures de la composition de la non-réponse. Naturellement, d'autres indicateurs R peuvent être construits en choisissant simplement différentes fonctions de distance entre les vecteurs des propositions à répondre. L'indicateur R et les graphiques présentés dans l'article peuvent être calculés en utilisant la plupart des logiciels statistiques standard.

Le calcul des indicateurs R est fondé sur un échantillon et s'appuie sur des modèles pour les propositions à répondre individuelles. Donc, ces indicateurs sont eux-mêmes des variables aléatoires et leur estimation comporte deux étapes qui influencent leur biais et leur variance. Cependant, c'est surtout la modélisation des propositions à répondre qui a des conséquences importantes. Le fait de se limiter à l'échantillon pour estimer les indicateurs R implique que ceux-ci sont moins précis, mais asymptotiquement, cette restriction n'introduit pas de biais. Le choix et l'ajustement du modèle sont habituellement effectués en imposant un niveau de signification et en ajoutant uniquement dans le modèle les interactions dont la contribution est significative. Ce dernier point signifie que la taille de l'échantillon et la disponibilité de données sur les variables auxiliaires jouent un rôle

Nous avons appliqué l'indicateur R proposé à deux études réalisées à Statistique Pays-Bas ces dernières années et dont les résultats avaient été examinés en détail par d'autres auteurs. L'accroissement ou la diminution de la valeur de l'indicateur R concorde avec les résultats des analyses plus détaillées effectuées par ces auteurs. Par conséquent, nous concluons que les indicateurs R peuvent être des outils valables. Cependant, un plus grand nombre de données empiriques sont manifestement nécessaires.

L'application de l'indicateur R a révélé qu'il n'existe aucune relation claire entre le taux de réponse et la représentativité de la réponse. Les taux de réponse plus élevés ne mènent pas nécessairement à une réponse plus équilibrée. Naturellement, nous constatons que les taux de réponse plus élevés réduisent le risque d'un biais de non-réponse. Plus le taux de réponse est élevé, plus le biais absolu maximal observé pour les items d'intérêt est faible.

L'application aux études choisies montre que les erreurs-types diminuent quand la taille de l'échantillon augmente, comme il fallait s'y attendre, mais elles demeurent relativement grandes pour des tailles d'échantillon modestes. Par exemple, pour une taille d'échantillon de 3 600, nous observons une erreur-type d'environ 1,3 %. Donc, si nous émettons l'hypothèse d'une loi normale, l'intervalle de confiance à 95 % a une largeur approximative de 5,4 %. La taille de l'échantillon de l'EPA est d'environ 30 000 unités. L'erreur-type est de l'ordre de 0,5 % et l'intervalle de confiance à 95 % correspondant à une largeur d'environ 2 %. Les erreurs-types sont plus grandes que nous ne l'avions pensé.

Le présent article contient une première étude empirique d'un indicateur R et de son erreur-type. Des travaux de recherche théoriques et empiriques beaucoup plus approfondis seront nécessaires pour comprendre pleinement les indicateurs R et leurs propriétés. En premier lieu, nous n'avons pas du tout tenu compte des items de l'enquête. Manifestement, il est absolument nécessaire que nous le fassions dans l'avenir. Cependant, comme nous l'avons déjà soutenu, les indicateurs R dépendent de l'ensemble de variables auxiliaires. Nous pouvons, par conséquent, conjecturer que, comme dans le cas des méthodes de correction de la non-réponse, la mesure dans laquelle les indicateurs R

IPAO autrement. Le tableau 6 donne les taux de réponse pour l'enquête ordinaire, la réponse à l'essai pilote pour l'enquête en ligne uniquement et la réponse à l'essai pilote dans son ensemble. La réponse à l'enquête en ligne uniquement est faible. Seulement 30 % de personnes ont rempli le questionnaire en ligne. Cela signifie que près de 70 % des unités échantillonnées ont été réaffectées à l'IPAO ou à l'ITAO. Cela a produit une réponse supplémentaire d'environ 35 %. Le taux global de réponse est légèrement plus faible que celui observé pour l'enquête ordinaire.

Fouwels, Janssen et Wetzeis (2006) ont procédé à une analyse univariée des compositions des réponses. Ils soutiennent que le taux de réponse est plus faible pour l'essai pilote, mais que cette diminution est relativement stable chez les divers sous-groupes démographiques. Ils constatent une baisse univariée du taux de réponse pour les variables auxiliaires d'âge, de groupe ethnique, de degré d'urbanisation et de type de ménage. Toutefois, ils relèvent des indices que la réponse devient moins représentative quand la comparaison est limitée aux personnes ayant répondu en ligne seulement. Cette constatation tient surtout pour l'âge des personnes échantillonnées, ce qui n'est pas étonnant.

Le tableau 6 donne la taille de l'échantillon, le taux de réponse, R , $CI_{0,05}^{RT}$, B_m et E_m pour trois groupes : l'enquête ordinaire, l'enquête pilote limitée au mode de collecte en ligne et l'enquête pilote dans son ensemble. Les variables auxiliaires d'âge, de groupe ethnique, de degré d'urbanisation et de type de ménage ont été appariées d'après des registres et ont été sélectionnées dans le modèle logistique pour les probabilités de réponse. Le tableau 6 montre que l'indicateur R est plus faible pour la réponse en ligne que pour l'enquête ordinaire. La valeur p correspondante est proche de 5 %. Étant donné à la fois le faible taux de réponse et le faible indicateur R , le biais absolu maximal B_m est plus de deux fois plus grand que pour l'enquête ordinaire. Par contre, pour l'essai pilote dans son ensemble, l'indicateur R ainsi que B_m sont proches des valeurs observées pour l'enquête ordinaire. À cause de la plus petite taille de l'échantillon de l'essai pilote, les erreurs-types estimées sont plus grandes que pour l'enquête ordinaire.

Réponse	n	Taux R	$CI_{0,05}^{RT}$	B_m	E_m
Ordinaire	30 139	68,9 %	81,4 %	(80,3-82,4)	6,8 %
Pilote – en ligne	3 615	30,2 %	77,8 %	(75,1-80,5)	18,3 %
Pilote – en ligne plus	3 615	64,7 %	81,2 %	(78,3-84,0)	7,3 %

Tableau 6
Taille de l'échantillon, taux de réponse, indicateur R , intervalle de confiance, biais maximal et $REQM$ maximale pour la réponse à l'enquête de surveillance de la sécurité ordinaire, l'enquête pilote avec composantes en ligne et suivi par IPAO/ITAO

Dans le tableau 4, nous notons une diminution de R quand nous comparons la réponse à l'EPA à la combinaison de cette réponse avec l'approche des questions de base. Cette diminution n'est pas significative. B_m diminue légèrement. Au tableau 5, cette comparaison est limitée aux ménages ayant un numéro de téléphone publié. En général, l'indicateur R est beaucoup plus élevé que pour l'ensemble des ménages. Comme la taille de l'échantillon est maintenant plus petite, les erreurs-types estimées sont plus grandes, comme en témoigne la largeur de l'intervalle de confiance. La valeur de B_m est plus faible. Pour la réponse combinée à l'EPA et à l'approche des questions de base, nous constatons un accroissement de R , mais de nouveau, il n'est pas significatif. B_m diminue.

Réponse	n	Taux R	$CI_{0,05}^{RT}$	B_m	E_m
EPA	10 135	68,5 %	86,3 %	(83,1-89,5)	5,0 %
EPA + questions de base	10 135	83,0 %	87,5 %	(84,3-90,7)	3,8 %

Dans le cas de l'exemple du suivi de l'EPA, nous constatons que les indicateurs R confirment les conclusions pour l'approche du rappel et celle des questions de base. En outre, l'accroissement de l'indicateur R consécutif à l'ajout de réponses au rappel est significatif au seuil de signification de 5 %.

5.3 Surveillance de la sécurité : essai pilote à mode mixte de 2006

En 2006, Statistique Pays-Bas a réalisé un essai pilote relatif à l'enquête de surveillance de la sécurité (*Safety Monitor*) en vue d'étudier les stratégies de collecte des données à mode mixte. Pour des renseignements détaillés, consulter Cobben, Janssen, Berkel et Brakel (2007). L'enquête de surveillance de la sécurité ordinaire, réalisée auprès de la population des Pays-Bas de 15 ans et plus, porte sur des questions relatives à la sécurité et à la façon dont la police remplit ses fonctions. Il s'agit d'une enquête à mode de collecte mixte. Les personnes dont le numéro de téléphone est publié sont approchées par ITAO. Celles qui ne peuvent pas être rejointes par téléphone sont approchées par IPAO. L'essai pilote de 2006 avait pour but d'évaluer la possibilité d'utiliser Internet comme l'un des modes dans une stratégie de collecte à mode mixte. Les personnes sélectionnées en ligne ont été approchées d'abord et participent au moyen d'une enquête en ligne. Les non-répondants à l'enquête en ligne ont été approchés de nouveau par ITAO s'ils possédaient un numéro de téléphone publié et par

estimons en utilisant l'échantillon, ce qui cause une perte de précision supplémentaire.

Un calcul analytique de l'erreur-type de R n'est pas simple à cause de l'estimation des probabilités de réponse.

Ici, nous résignons à utiliser des approximations numériques naïves de l'erreur-type. Nous estimons l'erreur-type de l'indicateur R par une méthode du bootstrap non paramétrique (Efron et Tibshirani 1993). Selon cette

méthode, l'erreur-type de l'indicateur R est estimée en sélectionnant un nombre $b = 1, 2, \dots, B$ d'échantillons bootstrap. Il s'agit d'échantillons tirés indépendamment et avec remise à partir de l'ensemble de données originales et

calculé pour chaque échantillon bootstrap b . Nous obtenons donc B répliques de l'indicateur R : R_{BT}^b , $b = 1, 2, \dots$

est une estimation de l'erreur-type de l'indicateur R , c'est-à-dire

$$s_{BT}^R = \sqrt{\frac{1}{B} \sum_{b=1}^B (R_{BT}^b - \hat{R}_{BT})^2} \quad (23)$$

où $\hat{R}_{BT} = 1/B \sum_{b=1}^B R_{BT}^b$ est l'indicateur R moyen estimé.

Dans les approximations, nous prenons $B = 200$ pour toutes les études. Nous avons expérimenté de plus grandes valeurs de B allant jusqu'à $B = 500$, mais avons constaté

dans tous les cas que l'estimation de l'erreur-type avait convergé dès la valeur $B = 200$.

Nous déterminons les intervalles de confiance 100(1 - α) % en supposant que la loi de R est approximativement normale et en employant les erreurs-types estimées données

par (23), soit

$$CI_{BT}^a = (R \pm \hat{s}_{BT}^{-\alpha} \times s_{BT}^R) \quad (24)$$

où $\hat{s}_{BT}^{-\alpha}$ est le quantile $1 - \alpha$ de la loi normale standard.

5.2 Enquête sur la population active – Étude de suivi de 2005

De juillet à décembre 2005, Statistique Pays-Bas a réalisé un suivi à grande échelle des cas de non-réponse à l'Enquête sur la population active (EPA) des Pays-Bas. L'étude a consisté en une nouvelle tentative d'interview auprès de deux échantillons de non-répondants à l'EPA, selon l'approche du rappel (Hansen et Hurwitz 1946) dans un cas et selon l'approche des questions de base (Kersten et Bethlehem 1984) dans l'autre. Les échantillons étaient formés de ménages sélectionnés pour participer à l'EPA qui avaient refusé de participer, dont les réponses n'avaient pas été traitées ou avec lesquels il n'y avait pas eu prise de contact pour l'EPA de juillet à octobre. Pour concevoir

l'étude de suivi, nous avons tenu compte des recommandations émanant des études de Stoop (2005) et de Voogt (2004).

Les principales caractéristiques des approches du rappel et des questions de base appliquées à l'EPA sont présentées au tableau 2. Pour plus de précisions, nous renvoyons le

lecteur à Schouten (2007) et à Cobben et Schouten (2007). Dans l'approche du rappel, nous avons employé la question

naïve originale visant les ménages administré par IPAQ, tandis que dans l'approche des questions de base, nous

avons utilisé un questionnaire abrégé dans des conditions de mode de collecte mixte. Le plan à mode mixte comportait

une application en ligne, une application papier et une application IPAQ. Cette dernière a été utilisée pour tous les

ménages possédant un numéro de téléphone publié. Ceux n'ayant pas de numéro de téléphone publié ont reçu une

lettre de présentation envoyée à l'avance, un questionnaire imprimé et la procédure d'entrée en communication avec un

site Internet sécurisé contenant le questionnaire en ligne. Les

répondants étaient libres de remplir le questionnaire imprimé ou le questionnaire en ligne.

Tableau 2 Caractéristiques des deux approches de l'étude de suivi

Approche du rappel	Approche des questions de base
• Questionnaire de l'EPA auquel devaient répondre tous les membres du ménage par IPAQ.	• Questionnaire fortement condensé contenant les questions clés de l'EPA et auquel on peut répondre en une à trois minutes.
• 28 intervieweurs sélectionnés indépendamment, selon les antécédents, obtenaient les meilleurs résultats.	• Plan de collecte des données à mode mixte (Internet, questionnaire papier et TTAO).
• Intervieweur différent de celui qui avait reçu la non-réponse.	• Questionnaire rempli par une personne par ménage, selon la méthode du prochain anniversaire.
• Formation supplémentaire concernant l'interaction au pas de la porte offerte aux intervieweurs.	• Ménage contacté une semaine après qu'il ait été traité comme un cas de non-réponse.
• Période de travail sur le terrain étendue de deux mois.	
• Intervieweur autorisé à offrir des mesures d'incitation.	
• Intervieweur susceptible de recevoir une prime.	
• Résumé imprimé des caractéristiques du ménage non répondant envoyé à l'intervieweur.	
• Affectation de l'adresse une semaine après la non-réponse.	

La taille de l'échantillon de l'enquête pilote de l'EPA était de $n = 18\,074$ ménages, parmi lesquels 11 275 ont répondu. Les ménages non répondants ont été stratifiés en fonction de la cause de la non-réponse. Étaient admissibles au suivi les ménages qui n'avaient pas été traités, ceux avec

nous arrivons à

$$\hat{R} \geq 1 - 4\hat{\sigma} \sqrt{y^2 - s^2(\hat{y}_{HT})} = r_2(y, \hat{p}). \quad (22)$$

Dans (20) et (22), nous représentons par $r_1(y, \hat{p})$ et $r_2(y, \hat{p})$ les bornes inférieures de l'indicateur R . Dans la section qui suit, nous donnons à $r_1(y, \hat{p})$ et $r_2(y, \hat{p})$ le nom de fonctions de représentativité de la réponse. Nous les calculons pour les études décrites à la section 5.

4.3.3 Exemple

Nous illustrons de nouveau la normalisation au moyen de l'exemple utilisé aux sections 2.1 et 3.3. La figure 3 contient la fonction de représentativité de la réponse $r_1(y, \hat{p})$ et les indicateurs R observés \hat{R} pour les six tentatives de prise de contact durant l'enquête POLS de 1998. Nous avons choisi trois valeurs de y , soit $y = 0,1$; $y = 0,075$ et $y = 0,05$.

La figure 3 indique qu'après la deuxième tentative de prise de contact, les valeurs de l'indicateur R sont supérieures à la borne inférieure correspondant au niveau de 10 %. Après quatre tentatives, l'indicateur R est proche du niveau de 7,5 %. Cependant, les valeurs ne sont jamais supérieures à l'autre borne inférieure qui est fondée sur le niveau de 5 %.

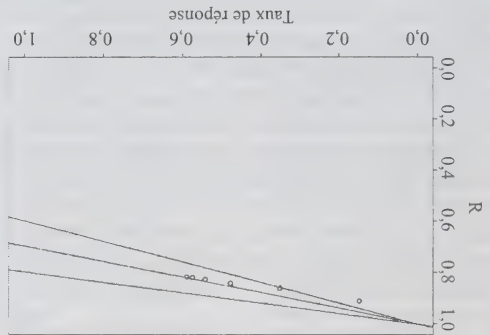


Figure 3 Bornes inférieures de l'indicateur R pour les six premières tentatives de prise de contact durant l'enquête POLS de 1998. Les bornes inférieures sont basées sur $y = 0,1$, $y = 0,075$ et $y = 0,05$

À la figure 4, le biais absolu maximal $\hat{B}_m(\hat{p})$ est représenté en fonction du taux de réponse pour les six tentatives de prise de contact. Après la troisième tentative, l'indicateur R a convergé vers une valeur autour de 8 %.

5. Application de l'indicateur R

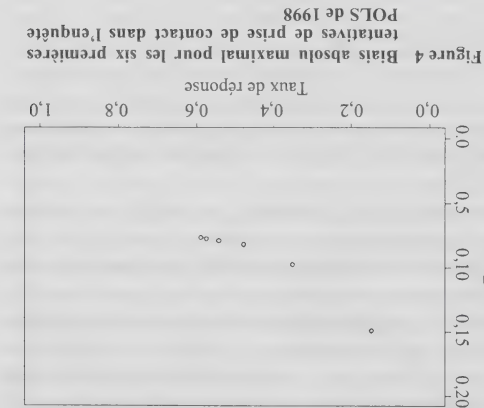


Figure 4 Biais absolu maximal pour les six premières tentatives de prise de contact dans l'enquête POLS de 1998

À la présente section, nous appliquons l'indicateur R à des études portant sur diverses stratégies de suivi des cas de non-réponse et diverses combinaisons de modes de collecte des données. La première a trait à l'Enquête sur la population active (EPA) des Pays-Bas. Elle comprend un examen de l'approche du rappel (Hansen et Hurwitz 1946) et de celle des questions de bases (Kersten et Bethlehem 1984). La deuxième concerne les plans de collecte de données à mode mixte appliqués dans le cadre de l'enquête nationale de surveillance de la sécurité (*National Safety Monitor*) des Pays-Bas.

Aux sections 5.2 et 5.3, nous examinons ces études de plus près en ce qui a trait à la représentativité de leurs différentes stratégies de travail sur le terrain. Avant cela, la section 5.1, nous décrivons comment nous obtenons une approximation des erreurs-types.

5.1 Erreur-type et intervalle de confiance

Si nous voulons comparer les valeurs de l'indicateur R pour diverses enquêtes ou stratégies de collecte des données, nous devons estimer leurs erreurs-types.

L'indicateur R fait intervenir l'écart-type d'échantillon des processus aléatoires entrant en jeu. Le premier est un échantillonage de la population, et le second, le mécanisme de réponse des unités échantillonnées. Si nous connaissons les probabilités de réponse réelles, le tirage d'un échantillon introduirait quand même une incertitude au sujet de l'indicateur R de population, et donc, entraînerait une certaine perte de précision. Comme nous ne connaissons pas les probabilités de réponse réelles, nous les

De nouveau, nous ne connaissons pas $E_m(p, y)$. À sa place, nous utilisons l'estimateur fondé sur l'échantillon qui emploie les probabilités de réponse estimées, désignées par $E_m(\hat{p}, y)$. Les bornes $B_m(\hat{p}, y)$ et $E_m(\hat{p}, y)$ sont différentes pour chaque item y de l'enquête. Pour les comparaisons, il est par conséquent commode de définir un item d'enquête hypothétique. Nous supposons que $S(y) = 0,5$. Nous désignons les bornes correspondantes par $B_m(\hat{p})$ et $E_m(\hat{p})$. Elles sont égales à

$$B_m(\hat{p}) = \frac{4\hat{p}}{(1 - R(\hat{p}))} \tag{17}$$

$$E_m(\hat{p}) = \sqrt{B_m^2(\hat{p}) + s^2(\hat{y}_{HT})}. \tag{18}$$

Nous calculons (17) et (18) dans le cas de toutes les études décrites à la section 5. Nous devons souligner que (17) et (18) sont de nouveaux des variables aléatoires qui possèdent une certaine précision et qui sont susceptibles de présenter un biais.

4.3.2 Fonctions de représentativité de la réponse

À la section précédente, nous avons utilisé l'indicateur R pour déterminer les bornes supérieures du biais de non-réponse et de la racine carrée de l'erreur quadratique moyenne de la moyenne (ajustée) des réponses. Inversement, nous pouvons établir une borne inférieure de l'indicateur R en demandant que le biais absolu de non-réponse ou la racine carrée de l'erreur quadratique moyenne soit inférieur à une valeur prescrite. Cette borne inférieure peut être choisie comme étant l'un des ingrédients des contraintes de qualité imposées aux données d'enquête par l'utilisateur de ces données. Si ce dernier ne veut pas que le biais de non-réponse ou la racine carrée de l'erreur quadratique moyenne excède une certaine valeur, l'indicateur R doit être supérieur à la borne correspondante. Il est évident que les bornes inférieures de l'indicateur R dépendent de l'item d'enquête. Par conséquent, nous nous limitons de nouveau à un item hypothétique pour lequel $S(y) = 0,5$.

$$B_m(\hat{p}) \leq y, \tag{19}$$

$$R \geq 1 - 4\hat{p} \hat{y} = r_1(y, \hat{p}). \tag{20}$$

$$E_m(\hat{p}) \leq y, \tag{21}$$

De manière analogue, en utilisant (18) et en demandant que

évidence, les problèmes d'interprétation soulevés à la section précédente ont également une incidence sur la normalisation de l'indicateur R . Par conséquent, ici, nous supposons que nous nous trouvons dans la situation idéale où nous pouvons estimer les propensions à répondre sans biais. Cette hypothèse tient pour les grandes enquêtes. Nous discutons la normalisation de l'indicateur R correspondant à R .

4.3.1 Biais absolu maximal et racine carrée de l'erreur quadratique moyenne

Nous montrons que, pour tout item d'une enquête Y , l'indicateur R peut être utilisé pour imposer des bornes supérieures au biais de non-réponse et à la racine carrée de l'erreur quadratique moyenne (REQM) des moyennes ajustées des réponses. Nous appliquons ces bornes à l'indicateur R pour montrer l'effet sous les pires scénarios. Soit Y une variable qui est mesurée dans une enquête et soit \hat{y}_{HT} l'estimateur Horvitz-Thompson de la moyenne de la population basée sur les réponses à l'enquête. On peut montrer (par exemple, Bethlehem 1988, Sanddal et Lundström 2005) que son biais $B(\hat{y}_{HT})$ est approximativement égal à

$$B(\hat{y}_{HT}) = \frac{C(y, p)}{C(y, \hat{p})}, \tag{14}$$

avec $C(y, p) = 1/N \sum_{i=1}^N (y_i - \bar{y})(p_i - \bar{p})$ la covariance de population entre les items étudiés et les probabilités de réponse. Pour une approximation étroite de la variance $s^2(\hat{y}_{HT})$ de \hat{y}_{HT} , nous nous référons à Bethlehem (1988).

L'inégalité de Cauchy-Schwarz permet de trouver une normalisation de R . Cette inégalité énonce que la covariance absolue par le produit des écarts-types des deux variables entre deux variables quelconques est bornée au sens absolu par le produit des écarts-types des deux variables. Nous pouvons traduire cela en bornes pour le biais (14) de \hat{y}_{HT}

$$|B(\hat{y}_{HT})| \leq \frac{\bar{p}}{S(\hat{p})S(y)} = \frac{1}{(1 - R(\hat{p}))S(y)} \frac{\bar{p}}{2\bar{p}}$$

$$= B_m(\hat{p}, y). \tag{15}$$

Manifestement, nous ne connaissons pas la borne supérieure $B_m(\hat{p}, y)$ dans (15), mais nous pouvons l'estimer en utilisant l'échantillon et les probabilités de réponse estimées. Nous désignons l'estimateur par $B_m(\hat{p}, y)$. De la même manière, nous pouvons établir une borne pour la racine carrée de l'erreur quadratique moyenne (REQM) de \hat{y}_{HT} . Il est vérifié approximativement que

$$\sqrt{B_m^2(\hat{p}, y) + s^2(\hat{y}_{HT})} \leq \sqrt{B_m^2(\hat{p}, y) + s^2(\hat{y}_{HT})} = E_m(\hat{p}, y). \tag{16}$$

L'échantillon, c'est-à-dire qu'il peut présenter un biais et qu'il possède une certaine exactitude. Mais qu'estime-t-il ? Supposons d'abord que la taille d'échantillon est arbitrairement grande de sorte que la précision ne joue aucun rôle et que la sélection d'un modèle pour les proportions à répondre ne pose pas de problème. Autrement dit, nous sommes capables d'ajuster tout modèle à tout ensemble fixe de variables auxiliaires.

Il existe une forte relation entre l'indicateur R , d'une part, et la disponibilité et l'utilisation de variables auxiliaires, d'autre part. À la section 2, nous avons défini la représentativité forte et faible. Même dans le cas où nous arrivons à ajuster n'importe quel modèle, nous ne sommes pas capables d'estimer les proportions à répondre au-delà du « pouvoir de résolution » des variables auxiliaires disponibles. Donc, nous ne pouvons tirer des conclusions qu'au sujet de la représentativité faible par rapport à l'ensemble de variables auxiliaires. Cela veut dire que, chaque fois que nous utilisons un indicateur R , nous devons utiliser comme complètement de sa valeur l'ensemble de covariables qui ont servi de grille pour estimer les proportions à répondre individuelles. Si l'indicateur R est utile pour des comparaisons, les ensembles de variables doivent être les mêmes. Il convient de souligner qu'il n'est pas nécessaire d'utiliser toutes les variables auxiliaires pour l'estimation des proportions à répondre, puisqu'elles peuvent ne pas accrotre la puissance explicative du modèle. Cependant, des ensembles identiques devraient être disponibles. L'indicateur R mesure alors un écart par rapport à la représentativité faible.

L'indicateur R ne relie pas les différences entre les probabilités de réponse à l'intérieur d'autres sous-groupes de la population que ceux définis par les classes de X . Si nous désignons de nouveau par $h = 1, 2, \dots, H$ les strates définies par X , par N_h la taille de la strate h et par p_h la moyenne de population des probabilités de réponse dans la strate h , il n'est pas difficile de montrer que R est une estimateur convergent de

$$R_X(p) = 1 - 2 \sqrt{\frac{1}{N-1} \sum_{h=1}^H N_h (p_h - \bar{p})^2}, \quad (13)$$

quand des modèles standard, comme la régression logistique ou la régression linéaire, sont utilisés pour estimer les probabilités de réponse. Naturellement, les indicateurs (13) et (5) peuvent être différents.

En pratique, la taille d'échantillon n'est pas arbitrairement grande. Elle a une incidence sur les deux étapes d'estimation, c'est-à-dire l'estimation des proportions à répondre et celle de l'indicateur R en utilisant un échantillon. Si nous connaissons les proportions à répondre individuelles, l'estimation fondée sur l'échantillon de l'indicateur R ne donnerait lieu qu'à une variance et ne produirait pas de biais. Nous serions capables d'estimer l'indicateur R de

population sans biais. Donc, pour les petites tailles d'échantillon, les estimateurs auraient une faible précision dont il pourrait être tenu compte en utilisant des intervalles de confiance au lieu de simples estimateurs ponctuels.

Les incidences sur l'estimation des probabilités de réponse sont toutefois différentes, à cause de la sélection et de l'ajustement du modèle. Il existe deux options. Nous pouvons imposer un modèle pour estimer les proportions à répondre en fixant les covariables au préalable ou nous pouvons laisser le modèle dépendre des covariables dont la contribution est significative par rapport à un niveau prédéfini. Dans le premier cas, nous n'introduirons de nouveau aucun biais, mais l'erreur-type pourrait être affectée par un surajustement. Dans le deuxième cas, le modèle pour l'estimation des proportions à répondre dépend de la taille de l'échantillon ; plus l'échantillon est grand, plus le nombre d'interactions acceptées comme étant significatives est élevé. Même si l'ajustement de modèles en se fondant sur un niveau de signification est une pratique statistique courante, la sélection du modèle peut introduire un biais et une variance dans l'estimation de tout indicateur R . Ce point est facile à comprendre si l'on considère la situation extrême d'un échantillon de taille 10. Pour un si petit échantillon, aucune interaction entre le comportement de réponse et les caractéristiques auxiliaires ne sera acceptée, ce qui laisse un modèle vide et un indicateur R estimé de 1. Les petits échantillons ne permettent tout simplement pas d'estimer les proportions à répondre. En général, un échantillon de petite taille conduira donc à une vue plus optimiste de la représentativité.

Nous devons faire ici une autre distinction subtile. Il se pourrait que, dans un sondage, un grand nombre d'interactions réunies contribuent à la prédiction des proportions à répondre, mais que les contributions individuelles soient très faibles, tandis que dans un autre sondage, une seule interaction intervienne, mais que sa contribution soit grande. Il se peut qu'individuellement, aucune petite contribution ne soit significative, mais qu'ensemble, elles soient aussi fortes que la grande contribution unique qui est significative. Donc, nous serions plus optimistes dans le premier exemple, même pour des tailles d'échantillon comparables.

Ces observations montrent qu'il faut toujours utiliser les indicateurs R avec prudence. Ils ne peuvent pas être considérés indépendamment des variables auxiliaires qui ont été utilisés pour les calculer. En outre, la taille de l'échantillon a une incidence sur le biais ainsi que sur la précision.

4.3 Normalisation

La troisième caractéristique importante d'un indicateur R est la normalisation. Nous voulons pouvoir donner des bornes à un indicateur R afin que son échelle, et donc, les variations de sa valeur, aient une signification. De toute

qui nous a permis de suivre la courbe de l'indicateur R durant la période de collecte des données. Pour l'estimation des taux de réponse, nous sommes servis d'un modèle de régression logistique comprenant la région, les antécédents ethniques et l'âge comme variables indépendantes. La région était une classification à 16 catégories, c'est-à-dire les 12 provinces et les quatre villes les plus grandes (Amsterdam, Rotterdam, La Haye et Utrecht) comme catégories distinctes. Les antécédents ethniques étaient répartis en sept catégories : natif, Marocain, Turc, Surinamais, Néerlandais, autre non-natif non occidental et autre non-natif occidental. La classification est fondée sur le pays de naissance des parents de la personne sélectionnée. La variable d'âge comprend trois catégories : de 0 à 34 ans, de 35 à 54 ans et de 55 ans et plus.

À la figure 2, R est représenté graphiquement en fonction du taux de réponse pour les six premières tentatives de prise de contact durant l'enquête POLS. La valeur la plus à gauche correspond à l'ensemble de répondants après une tentative. Pour chaque tentative supplémentaire, le taux de réponse augmente, mais l'indicateur révèle une diminution de la représentativité. Ce résultat confirme les constatations faites par Schouten (2004).

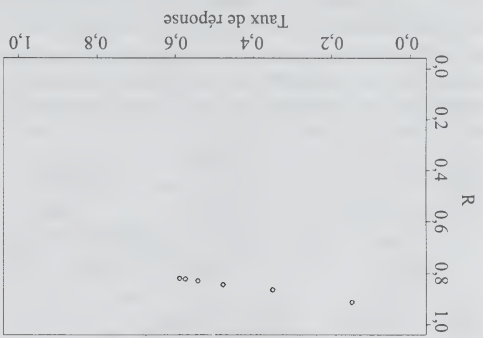


Figure 2 Indicateur R pour les six premières tentatives de prise de contact dans l'enquête POLS de 1998

4. Caractéristiques des indicateurs R

À la section 3, nous proposons un indicateur de la représentativité. Cependant, il est possible d'en construire d'autres. Les mesures d'association ou les indices d'adéquation sont nombreux, par exemple, Goodman et Kruskal (1979), Bentler (1990) et Marsh, Balla et McDonald (1988). La relation entre les mesures d'association et les indicateurs R est forte. Essentiellement, ces indicateurs ont pour objectif

4.1 Caractéristiques générales

Nous voulons que les indicateurs R soient fondés sur une fonction ou une mesure de distance au sens mathématique. La propriété d'irrégularité triangulaire d'une fonction de distance permet un classement partiel de la variation dans les propensions à répondre qui facilite l'interprétation. Une fonction de distance peut être dérivée facilement de toute norme mathématique. À la section 3, nous choisissons la norme euclidienne qui est celle utilisée couramment. Elle nous mène à un indicateur R qui s'appuie sur l'écart-type des propensions à répondre. D'autres normes, comme la norme supremum, ou norme sup, nous mèneraient à d'autres fonctions de distance. Toutefois, à la section 4.3, nous montrons que les indicateurs R fondés sur la norme euclidienne ont des caractéristiques de normalisation intéressantes.

Nous devons faire une distinction subtile entre les indicateurs R et les fonctions de distance. Ces dernières sont symétriques, tandis qu'un indicateur R mesure un écart par rapport à un point particulier, à savoir celui où toutes les propensions à répondre sont égales. Si nous modifions le vecteur des propensions individuelles, dans la plupart des cas, ce point se déplace. Toutefois, si nous fixons la valeur de la propension à répondre moyenne, la fonction de distance facilite l'interprétation.

À part une relation avec une fonction de distance, nous voulons pouvoir mesurer, interpréter et normaliser les indicateurs R. À la section 3.2, nous avons déjà dérivé des estimateurs fondés sur la réponse pour les indicateurs R « de population » qui ne sont pas mesurables quand les propensions à répondre sont inconnues et que l'on ne dispose que des réponses à un sondage. Donc, nous avons rendu les indicateurs R mesurables en passant à des estimateurs. Les deux autres caractéristiques sont examinées séparément dans les deux sections qui suivent.

4.2 Interprétation

La deuxième caractéristique des indicateurs R est la facilité avec laquelle nous pouvons interpréter leur valeur et la comparaison des indicateurs R. Nous sommes passés à un estimateur pour un indicateur R qui est fondé sur les échantillons des sondages et sur des estimateurs de probabilité de réponse individuelle. Ces deux éléments ont des conséquences très importantes en ce qui concerne l'interprétation et la comparaison des indicateurs R. Puisque l'indicateur R est lui-même un estimateur, il dépend de

l'ensemble des classes h , et que nous supposons que les variances intraclasses sont nulles, il est vérifié que

$$S^2(p) = \frac{1}{H} \sum_{h=1}^H N_h (p_h - \bar{p})^2$$

$$= \frac{N}{H} \sum_{h=1}^H f_h(p_h - \bar{p})^2 \approx \sum_{h=1}^H f_h(p_h - \bar{p})^2. \quad (7)$$

La statistique χ^2 mesure la distance entre les proportions observées et prévues. Cependant, il ne s'agit d'une fonction de distance varie qu'au sens mathématique pour les distributions marginales fixes f_h et \bar{p} . Nous pouvons appliquer la statistique χ^2 à X afin de « mesurer » la distance entre le comportement de réponse réel et le comportement de réponse qui est attendu quand la réponse est indépendante de X . Autrement dit, nous mesurons l'écart par rapport à la représentativité faible en ce qui concerne X .

Nous pouvons réécrire la statistique χ^2 pour obtenir

$$\chi^2 = \sum_{h=1}^H \frac{N_h \bar{p}}{(N_h p_h - N_h \bar{p})^2} + \sum_{h=1}^H \frac{N_h (1 - \bar{p})}{(N_h (1 - p_h) - N_h (1 - \bar{p}))^2}$$

$$= \sum_{h=1}^H \frac{N_h \bar{p}}{N_h^2 (p_h - \bar{p})^2} + \sum_{h=1}^H \frac{p}{N_h^2 (1 - p_h - (1 - \bar{p}))^2}$$

$$= \frac{N}{H} \sum_{h=1}^H f_h(p_h - \bar{p})^2$$

$$= \frac{N}{H} \sum_{h=1}^H f_h(p_h - \bar{p})^2$$

(8)

Une mesure d'association qui converti la statistique χ^2 à l'intervalle $[0, 1]$ (voir, par exemple, Agresti 2002) est le V de Cramér

$$V = \sqrt{\frac{\chi^2}{N(\min\{C, R\} - 1)}} \quad (9)$$

où C et R sont, respectivement, le nombre de colonnes et de lignes dans la table de contingence sous-jacente. Le V de Cramér atteint une valeur 0 si les proportions observées concordent exactement avec les proportions espérées et sa valeur maximale est 1. Dans notre cas, le dénominateur est égal à N puisque l'indicateur de réponse ne possède que deux catégories : réponse et non-réponse. Par conséquent, (9) se transforme en

$$V = \sqrt{\frac{\chi^2}{N - 1}} = \sqrt{\frac{N p (1 - p)}{N - 1}} S(p). \quad (10)$$

Partant de (10), nous voyons que, si N est grand, le V de Cramér est approximativement égal à l'écart-type des proportions à répondre normalisées par l'écart-type maximal $\sqrt{p(1 - p)}$ pour une propension à répondre moyenne fixe \bar{p} .

3.2 Indicateurs R fondés sur la réponse

À la section 3.1, nous avons supposé que les propensions à répondre individuelles étaient connues. Naturellement, en pratique, elles ne le sont pas. En outre, dans un sondage, nous ne possédons des renseignements que sur le comportement de réponse des unités échantillonnées. Par conséquent, nous devons trouver d'autres solutions pour les indicateurs R . Un moyen évident consiste à utiliser des estimateurs fondés sur la réponse pour les propensions à répondre individuelles et la propension à répondre moyenne. Soit \hat{p}_i un estimateur de p_i qui utilise toutes les variables auxiliaires disponibles ou un sous-ensemble de celles-ci. Les méthodes qui permettent de calculer ce genre d'estimation sont, par exemple, les modèles de régression logit ou probit (Agresti 2002) et les arbres de classification CHAID (Kass 1980). Soit $\hat{\bar{p}}$ la moyenne pondérée d'échantillon des propensions à répondre estimées, c'est-à-dire

$$\hat{\bar{p}} = \frac{1}{N} \sum_{i=1}^N \hat{p}_i \pi_i, \quad (11)$$

où nous utilisons les poids d'inclusion. Nous remplaçons R par l'estimateur \hat{R}

$$\hat{R}(p) = 1 - 2 \sqrt{\frac{1}{N} \sum_{i=1}^N \pi_i (\hat{p}_i - \hat{\bar{p}})^2}. \quad (12)$$

Notons que, dans (12), il existe en fait deux étapes d'estimation fondées sur des mécanismes probabilistes différents. Les propensions à répondre proprement dites sont estimées, ainsi que la variation dans les propensions. Nous revenons sur les conséquences de l'estimation en deux étapes à la section 4.

3.3 Exemple

Nous appliquons les indicateurs R proposés à des données provenant de l'enquête POLS de 1998 décrite à la section 2.1. Rappelons que l'enquête consistait en une combinaison d'interviews sur place et d'interviews téléphoniques dans laquelle les interviews du premier mois étaient de type PPAO uniquement. La taille de l'échantillon était de près de 40 000 et le taux de réponse était de 60 % environ. Nous avons relié des données administratives concernant le travail sur le terrain à l'échantillon et déduit si chaque tentative de prise de contact avait abouti à une réponse, ce

$$\bar{p}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} p_{hk} = p, \text{ pour } h = 1, 2, \dots, H, \quad (2)$$

où N_h est la taille de la population de la catégorie h , p_{hk} est la propension à répondre de l'unité k dans la classe h et la somme est faite sur toutes les unités dans cette catégorie.

La définition faible correspond à un mécanisme de création de données manquantes de type MCAR par rapport à X , car selon l'hypothèse MCAR, nous ne pouvons pas faire la distinction entre les répondants et les non-répondants en nous fondant sur la connaissance de X .

3. Indicateurs R

À la section précédente, nous avons défini une réponse fortement représentative et faiblement représentative. Les deux définitions s'appuient sur les probabilités de réponse individuelles qui sont inconnues en pratique. Pour commencer, nous prenons un indicateur R de population. Puis, les propensions à répondre sont estimées.

3.1 Indicateurs R de population

Nous considérons d'abord la situation hypothétique où les propensions à répondre individuelles sont connues. Manifestement, dans ce cas, nous pouvons même tester la définition forte et nous cherchons simplement à mesurer la quantité de variation dans les propensions à répondre ; la représentativité au sens fort est d'autant moindre que la variation est forte. Soit $p = (p_1, p_2, \dots, p_N)'$ un vecteur de propensions à répondre, soit $\mathbf{1} = (1, 1, \dots, 1)'$ le vecteur de dimension N de valeurs 1, et soit $p_0 = \mathbf{1} \times \bar{p}$ le vecteur constitué de la propension à répondre moyenne de la population. Toute fonction de distance d dans $[0, 1]^N$ suffirait à mesurer l'écart par rapport à une réponse fortement représentative en calculant $d(p, p_0)$. Notons que la hauteur de la réponse globale ne joue aucun rôle. La distance euclidienne est une fonction de distance simple. Appliquée à une distance entre p et p_0 , cette mesure est proportionnelle à l'écart-type des probabilités de réponse

$$S(p) = \sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - \bar{p})^2}. \quad (3)$$

Il n'est pas difficile de montrer que

$$S(p) \leq \sqrt{\bar{p}(1 - \bar{p})} \leq \frac{1}{2}. \quad (4)$$

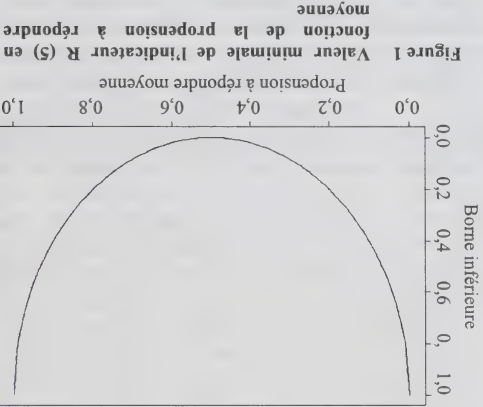


Figure 1 Valeur minimale de l'indicateur R (S) en fonction de la propension à répondre moyenne

Notons que la valeur minimale de (5) dépend du taux de réponse (voir la figure 1). Pour $\bar{p} = 1/2$, il possède une valeur minimale de 0. Pour $\bar{p} = 0$ et $\bar{p} = 1$, aucune variation n'est manifestement possible et la valeur minimale est 1. Paradoxalement, la borne inférieure augmente quand le taux de réponse diminue pour passer de $1/2$ à 0. Dans le cas d'un taux de réponse faible, la possibilité que la variation des propensions à répondre individuelles soit forte est moindre.

$$R(p) = 1 - 2S(p). \quad (5)$$

Nous voulons que l'indicateur R prenne les valeurs comprises dans l'intervalle $[0, 1]$, la valeur 1 représentant une représentativité forte et la valeur 0 étant l'écart maximal par rapport à la représentativité forte. Nous proposons l'indicateur R , que nous désignons par R , défini par

$$f_h = \frac{N}{n_h}, \text{ pour } h = 1, 2, \dots, H. \quad (6)$$

Nous pouvons considérer R comme étant une mesure du manque d'association. Quand $R(p) = 1$, il n'existe aucune relation entre tout item d'une enquête et le mécanisme de création de données manquantes. Nous montrons que R est souvent utilisée pour tester l'indépendance et l'adéquation. Supposons que les propensions à répondre diffèrent uniquement pour les classes h définies par une variable catégorique X . Soit \bar{p}_h et f_h , respectivement, la propension à répondre et la fonction de population de la classe h , c'est-à-dire

Donc, pour tout i avec $X_i = h$, la propension à répondre est $p_i = \bar{p}_h$. Puisque la variance des propensions à répondre est la somme des variances « interclasses » et « intraclasses » sur

Le concept de réponse représentative est aussi relié étroitement aux mécanismes de création des données manquantes désignés MCAR (pour *Missing-Completely-at-Random* ou Manquant entièrement au hasard), MAR (pour *Missing-at-Random* ou Manquant au hasard) et NMAR (pour *Not-Missing-at-Random* ou Ne manquant pas au hasard) qui sont souvent mentionnés dans la littérature ; voir Little et Rubin (2002). Un mécanisme de création de données manquantes est de type MCAR quand la probabilité de réponse ne dépend pas du sujet d'enquête d'intérêt. Le mécanisme est de type MAR si la probabilité de réponse dépend des données observées seulement, ce qui est donc une hypothèse plus faible que l'hypothèse MCAR. Si la probabilité dépend également des données manquantes, le mécanisme est de type NMAR. En fait, ces mécanismes ont pour origine la théorie statistique fondée sur la modélisation. Interprété plus ou moins librement en ce qui a trait à un sujet d'enquête, MCAR signifie que les répondants sont, en moyenne, les mêmes que les non-répondants, MAR signifie que, au sein de sous-populations connues, les répondants sont en moyenne les mêmes que les non-répondants, et NMAR implique que, même au sein de sous-populations, les répondants sont différents des non-répondants. L'ajout du sujet d'enquête est essentiel. Dans un même questionnaire, certains items peuvent être de type MCAR, tandis que d'autres sont de type MAR ou NMAR. En outre, l'hypothèse d'un mécanisme MAR pour un item tient pour une stratification particulière de la population. Un autre item pourrait nécessiter une stratification différente.

Étant donné que nous désirons surveiller et comparer la réponse à des enquêtes qui diffèrent par le sujet étudié ou par la période où elles sont exécutées, il n'est pas intéressant de définir une réponse représentative comme étant dépendante du sujet d'enquête proprement dit ni comme étant concentrée sur la qualité de la collecte des données plutôt que sur l'estimation. Ces conditions nous amènent à comparer la composition de la réponse (groupe de répondants) à celle de l'échantillon. Clairement, les sujets étudiés influencent la probabilité que les ménages participent à l'enquête, mais cette influence ne peut pas être mesurée ni testée et, par conséquent, de notre point de vue, ils ne peuvent pas représenter les données d'entrée utilisées pour évaluer la qualité de la réponse. Nous proposons de juger la composition de la réponse au moyen d'ensembles prédéfinis de variables qui sont observés en dehors du cadre de l'enquête et qui peuvent être employés pour chaque enquête examinée. Nous voulons que la sélection des répondants soit aussi proche que possible d'un « échantillon aléatoire simple de l'échantillon utilisé pour l'enquête », c'est-à-dire présentant une relation aussi faible que possible entre la réponse et les caractéristiques qui distinguent les unités les

plus des autres. Le dernier énoncé peut être interprété comme signifiant l'absence de forces sélectives dans la sélection des répondants, ou comme un mécanisme de type MCAR en ce qui concerne toutes les variables étudiées possibles.

2.2 Définition d'un sous-ensemble de réponses représentatif

Soit $i = 1, 2, 3, \dots, N$ les étiquettes d'unité pour la population. Nous désignons par s_i l'indicateur d'échantillon 0-1, c'est-à-dire que l'unité i prend la valeur 1 si elle est échantillonnée, et 0 autrement. Nous désignons par η_i l'indicateur de réponse 0-1 pour l'unité i . Si l'unité i est échantillonnée et qu'elle répond, alors $\eta_i = 1$. Sinon, sa valeur est 0. La taille d'échantillon est n . Enfin, η_i^* désigne la probabilité d'inclusion de premier ordre de l'unité i . La clé de nos définitions réside dans les proportions à répondre quand elle est échantillonnée.

L'interprétation d'une propension à répondre n'est pas directe. Nous suivons une approche assistée par modèle, ce qui signifie que seuls l'échantillon et les indicateurs de réponse ont une caractéristique d'une unité étiquetée et identifiable, si l'on peut dire une pièce de monnaie biaisée que l'unité garde en poche, et est, par conséquent, inséparable de cette unité. Toutefois, moyennant un peu d'effort, tous les concepts peuvent être traduits dans un contexte fondé sur un modèle.

Pour commencer, nous donnons une définition forte.

Définition (forte) : Un sous-ensemble de réponses représentatif de l'échantillon si les proportions à répondre p_i sont les mêmes pour toutes les unités de la population

$$p_i = P[\eta_i = 1 \mid s_i = 1] = p, \quad \forall i, \quad (1)$$

et si la réponse d'une unité est indépendante de la réponse de toutes les autres unités.

Si un mécanisme de données manquantes satisfait la définition forte, il sera du type Manquant entièrement au hasard (MCAR) pour toutes les questions d'enquête possibles. Bien que la définition soit séduisante, il est impossible de tester sa validité en pratique. Nous ne possédons aucune répétition de la réponse d'une unité unique. Par conséquent, nous construisons aussi une définition faible qui peut être testée en pratique.

Définition (faible) : Un sous-ensemble de réponses est représentatif d'une variable catégorique X possédant H catégories si la propension à répondre moyenne sur les catégories est constante, soit

et les variables auxiliaires. En fait, nous considérons l'indicateur R comme une mesure du manque d'association. La situation est d'autant meilleure que l'association est faible, car cela signifie qu'il n'existe aucune preuve que la non-réponse a affecté la composition des données observées.

Afin de pouvoir utiliser les indicateurs R comme outils de surveillance et de comparaison de la qualité des enquêtes dans l'avenir, ceux-ci doivent avoir les caractéristiques d'une mesure. Autrement dit, nous voulons un indicateur R qui peut être interprété, mesuré et normalisé tout en ayant les propriétés mathématiques d'une mesure. Et ce, surtout parce que l'interprétation et la normalisation ne sont pas des caractéristiques directes.

Nous appliquons l'indicateur R à deux études menées à Statistiques Pays-Bas en 2005 et en 2006. Elles avaient pour objectifs de comparer différentes stratégies de collecte des données. Elles comportaient divers modes de collecte des données et diverses stratégies de suivi des cas de non-réponses. Chacune de ces études a été suivie d'une analyse détaillée des données et de la production de rapports. Par conséquent, elles conviennent bien pour une validation empirique de l'indicateur R. Nous comparons les valeurs de l'indicateur R aux conclusions des analyses. Nous revoions le lecteur à Schouten et Cobben (2007), ainsi qu'à Cobben et Schouten (2007) pour d'autres exemples et études empiriques.

À la section 2, nous commençons par discuter du concept de réponse représentative. Puis, à la section 3, nous définissons la notation mathématique pour notre indicateur. La section 4 est consacrée aux caractéristiques de l'indicateur R. La section 5 décrit l'application de cet indicateur R aux études sur le terrain. Enfin, à la section 6, nous présentons une discussion.

2. Le concept de réponse représentative

En premier lieu, nous expliquons ce que nous entendons quand nous disons qu'un ensemble de répondants à une enquête est représentatif de l'échantillon. Ensuite, nous rendons le concept de représentativité mathématiquement rigoureux en lui donnant une définition.

2.1 Que signifie représentatif ?

La littérature nous avertit de ne pas concentrer tous nos efforts sur les taux de réponse pour évaluer la qualité des enquêtes. Nous pouvons illustrer facilement ce point au moyen d'un exemple tiré de l'enquête néerlandaise POLS (acronyme de *Permanent Onderzoek Leefsituatie* ou Enquête intégrée sur les conditions de vie des ménages en français).

L'exemple semble montrer clairement que l'effort accru a abouti à une réponse moins représentative en ce qui a trait aux deux variables auxiliaires. Mais d'une manière générale, qu'entendons-nous par représentatif ?

Tableau 1
Moyennes des réponses à l'enquête POLS pour le premier mois de collecte et pour la période complète de collecte de deux mois

Variable	Après 1 mois	Après 2 mois	Echantillon
Recevant des allocations sociales	10,5 %	10,4 %	12,1 %
Parent non naif	12,9 %	12,5 %	15,0 %
Taux de réponse	47,2 %	59,7 %	100 %

Le tableau 1 contient les estimations, fondées sur les réponses à l'enquête POLS après un mois et après deux mois, de la proportion de la population néerlandaise qui recevait une forme d'allocations sociales et de la proportion dont au moins un parent était né ailleurs qu'aux Pays-Bas. Pour les deux variables, les données sont extraites de registres et sont traitées artificiellement comme des réponses à une enquête en supprimant leurs valeurs pour les non-répondants. Les proportions dans l'échantillon sont également données au tableau 1. Après un mois, le taux de réponses était de 47,2 %, tandis qu'après la période complète d'interview de deux mois, il était de 59,7 %. Dans le cas de l'enquête POLS de 1998, le premier mois, la collecte a été effectuée par IPAO (interview sur place assistée par ordinateur). Après le premier mois, les non-répondants ont été affectés à un mode de collecte par ITAO (interview téléphonique assistée par ordinateur) s'ils étaient abonnés à un service téléphonique par fil dont le numéro était publié. Sinon, ils ont de nouveau été affectés à un mode de collecte par IPAO. Donc, le deuxième mois d'interview a donné 12,5 % supplémentaires de réponses. Cependant, l'examen du tableau 1 montre qu'après le deuxième mois, les estimations fondées sur les données de l'enquête présentent un biais plus important qu'après le premier mois.

Indicateurs de la représentativité de la réponse aux enquêtes

Barry Schouten, Fannie Cobben et Jelte Bethlehem

Résumé

De nombreux organismes statistiques considèrent le taux de réponse comme étant l'indicateur de la qualité à utiliser en ce qui concerne l'effet du biais de non-réponse. Ils prennent donc diverses mesures en vue de réduire la non-réponse ou de maintenir la réponse à un niveau jugé acceptable. Cependant, à lui seul, le taux de réponse n'est pas un bon indicateur du biais de non-réponse. En général, un taux de réponse élevé n'implique pas que le biais dû à la non-réponse est faible. On trouve à cet égard de nombreux exemples dans la littérature (par exemple, Groves et Peytcheva 2006 ; Keeter, Miller, Kohut, Groves et Presser 2000 ; Schouten 2004).

Nous introduisons un certain nombre de concepts et un nouvel indicateur en vue d'évaluer la similitude entre la réponse à une enquête et l'échantillon de cette enquête. Cet indicateur de la qualité, que nous appelons indicateur R, peut servir de complément aux taux de réponse et est destiné principalement à évaluer le biais de non-réponse. Il peut faciliter l'analyse de la réponse aux enquêtes en fonction du temps, ou pour diverses stratégies d'enquête sur le terrain ou divers modes de collecte des données. Nous appliquons l'indicateur R à deux exemples pratiques.

Mots clés : Qualité ; non-réponse ; réduction de la non-réponse ; correction de la non-réponse.

1. Introduction

Il est bien décrit dans la littérature sur la méthodologie d'enquête que les taux de réponse sont, à eux seuls, de médiocres indicateurs du biais de non-réponse ; voir, par exemple, Curtin, Presser et Singer (2000), Groves, Presser et Dippo (2004), Groves (2006), Groves et Peytcheva (2006), Keeter et coll. (2000), Merkle et Edelman (2002), Heerwegh, Abis et Loosveldt (2007), ainsi que Schouten (2004). Cependant, les spécialistes du domaine n'ont pas encore établi d'autres indicateurs de la non-réponse qui seraient moins ambigus en tant qu'indicateurs de la qualité d'une enquête.

Nous proposons un indicateur, que nous appelons indicateur R (« R » pour représentativité), de la similitude entre la réponse à une enquête, d'une part, et l'échantillon ou la population à l'étude, d'autre part. Cette similitude peut être appelée « réponse représentative ». La littérature spécialisée offre de nombreuses interprétations du concept de « réponse représentative ». Consulter Kruskal et Mosteller (1979a, b et c) pour une étude approfondie de la littérature statistique et non statistique. Rubin (1976) a introduit le concept de non-réponse ignorable, c'est-à-dire les conditions minimales permettant d'obtenir une estimation sans biais d'une statistique. Certains auteurs définissent explicitement la représentativité. Hájek (1981) relie le terme « représentatif » à l'estimation de paramètres de population ; la paire formée par un estimateur et un mécanisme de création de données manquantes est représentative si, avec une probabilité de un, l'estimateur est égal au paramètre de population. Suivant la définition de Hájek, les estimateurs par calage (par exemple,

L'indicateur R que nous proposons s'appuie sur des probabilités de réponse estimées. L'estimation de ces probabilités implique que l'indicateur R proprement dit est une variable aléatoire et, par conséquent, qu'il possède une précision et éventuellement un biais. La taille de l'échantillon d'une enquête joue donc un rôle important dans l'évaluation de l'indicateur R, comme nous le montrerons. Cependant, cette dépendance existe pour toute mesure ; les petites enquêtes ne permettent tout simplement pas de tirer des conclusions catégoriques au sujet du mécanisme de création des données manquantes.

Nous montrons que l'indicateur R proposé est apparenté à la mesure V de Cramér pour l'association entre la réponse

Sämdal, Swenson et Wretman (2003) sont représentatifs pour les variables auxiliaires qui sont calées. Bertino (2006) définit un indice de représentativité dit univarié pour des variables aléatoires continues. Cet indice est une mesure d'exemple de distribution basée sur la statistique de Cramér-Von Mises. Kohler (2007) définit ce qu'il appelle un critère interne de représentativité. Ce critère univarié ressemble à la statistique Z pour les moyennes de population.

Nous isolons le concept de représentativité de l'estimation d'un paramètre de population particulier, mais nous le relient à l'effet sur la composition globale de la réponse. Dissocier les indicateurs d'un paramètre particulier permet de les utiliser comme outils pour comparer diverses enquêtes entre elles ou les résultats d'une enquête au cours du temps, ainsi que pour comparer des stratégies et modes différents de collecte des données. En outre, la mesure donne une perspective multidimensionnelle de la dissimilitude entre l'échantillon et la réponse.

- Chambers, R.L., et Skinner, C.J. (Eds.) (2003). *Analysis of Survey Data*. Wiley, Chichester.
- Collins, M., et Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.
- Davis, P.D., et Scott, A.J. (1995). La variance de l'intervieweur et ses effets sur les comparaisons de domaines. *Techniques d'enquête*, 21, 111-118.
- Feather, J. (1973). A study of interviewer variance. WHO International Collaborative Study of Medical Care Utilization, Saskatchewan Study Area Reports, Séries II, Monographie No. 3.
- Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Gabler, S., Häder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.
- Gray, P.G. (1956). Examples of interviewer variability taken from two sample surveys. *Applied Statistics*, V, 73-85.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.
- Groves, R.M., et Magliav, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.
- Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol I, II. New York : John Wiley & Sons, Inc.
- Hansen, M.H., Hurwitz, W.N. et Bershad, M.A. (1961). Measurement errors in census and surveys. *Bulletin of the ISI* 38, 2, 351-374.
- Hanson, R.H., et Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.
- Häček, J. (1971). Comments. In Foundations of Statistical Inference, (Eds. V.P. Godambe et D.A. Sprott). Toronto : Holt, Rinehart, and Winston.
- Heeb, J.-L., et Gmel, G. (2001). Interviewers and respondents effects on self-reported alcohol consumption in Swiss Health Survey. *Journal of Studies on Alcohol*, 62, 434-442.
- Hox, J.J., et De Leeuw, E.D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modeling to meta-analysis. *Quality & Quantity*, 329-344.
- Kalton, G., Brick, J.M. et Lè, T.H. (2005). Estimating components of design effects for use in sample design. Dans : *Household Sample Surveys in Developing and Transition Countries*. Chapitre VI. Disponible au http://unstats.un.org/unsd/hhsurveys/pdf/Chapter_6.pdf.
- Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.
- Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- Lynn, P., et Gabler, S. (2004). Approximations to b^* in the estimation of design effects due to clustering. *Working Papers of the Institute for Social and Economic Research*, papier 2004-07. Colchester : University of Essex. Disponible au <http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-07.pdf>.
- Malalalobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, Série A, 109, 325-378, ré-édité dans *Sankhyā* (1958), 1-68.
- O'Muircheartaigh, C., et Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society*, Séries A, 161, 63-77.
- Phillipens, M., et Loosveld, G. (2004). Interviewer-related variance in the European Social Survey. Article présenté à la sixième conférence internationale sur la méthodologie des sciences sociales, 17-20 août, Amsterdam.
- Rice, S.A. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology*, 35, 420-423.
- Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multi-way contingency tables with proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- Schnell, R., et Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.
- Scott, A.J., et Davis, P.D. (2001). Estimation des effets de l'intervieweur sur des réponses binaires. *Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique*.
- Skinner, C.J. (1986). Design effect of two-stage sampling. *Journal of the Royal Statistical Society*, Série B, 48, 89-99.
- Skinner, C.J., Holt, D. et Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York : John Wiley & Sons, Inc.
- Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47, 84-95.
- Valliant, R.M. (1987). Generalized variance functions in stratified two-stage sampling. 82, 499-508.
- Verma, V., Scott, C. et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Série A, 143, 431-473.

Maintenant, le résultat s'ensuit en effectuant certaines opérations algébriques.

Résultat 5.

$$eff_{s,w} = \frac{Var_{M_4}(\bar{y}_w)}{Var_{M_3}(\bar{y}_w)}$$

$$= \frac{\frac{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2} - 1}{\frac{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2} - 1} + p_c$$

Preuve. Le résultat s'obtient en notant que

$$\frac{Var_{M_4}(\bar{y}_w)}{Var_{M_3}(\bar{y}_w)} =$$

$$\frac{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2 + p_c \sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2 + p_c \sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2 + p_c \sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2 + p_c \sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}$$

et en effectuant certaines opérations algébriques.

Résultat 6. Pour $0 < p_c < 1$ et $0 < p_{int} < 1$,

$$eff \geq 1 + p_c \left(\frac{p}{n} - 1 \right) + p_{int} \left(\frac{I}{n} - 1 \right),$$

avec égalité si et uniquement si les poids sont tous égaux et que les intervieweurs ont tous la même charge de travail.

Si, dans chaque UPE, nous n'avons qu'un seul intervieweur, alors

$$eff \geq 1 + (p_c + p_{int}) \left(\frac{p}{n} - 1 \right).$$

Preuve : En utilisant certaines notions algébriques et l'inégalité générale,

$$\sum_j^f p_j x_j^2 \geq \left(\sum_j^f p_j x_j \right)^2$$

avec

$$p_j \geq 0 \text{ et } \sum_{j=1}^f p_j = 1,$$

nous avons

$$eff = \frac{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2} (1 - p_c - p_{int})$$

$$+ n p_c \frac{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2} + n p_{int} \frac{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2}{\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2} \geq 1 - p_c - p_{int} + p_c \frac{p}{n} + p_{int} n \left(\sum_{p=1}^P \sum_{l_p=1}^{I_p} \sum_{n_p=1}^{N_p} w_{pik}^2 \right)^{-1}$$

$$\geq 1 - p_c - p_{int} + p_c \frac{p}{n} + p_{int} \frac{I}{n} \left(\frac{p}{n} - 1 \right).$$

Remerciements

Nous remercions le rédacteur en chef et les examinateurs de leurs suggestions et commentaires constructifs, qui ont amélioré la version originale de cet article.

Bibliographie

- Bailar, B.A., Bailey, L. et Stevens, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.
- Bailey, L., Moore, T.F. et Bailar, B.A. (1978). An interviewer variance study for eight impact cities of the National Crime Survey Cities Sample. *Journal of the American Statistical Association*, 73, 16-23.
- Biemer, P., et Stokes, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 369, 158-166.

- Biemer, P., et Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. Dans *Survey Measurement and Process Quality*, (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York: John Wiley & Sons, Inc., 603-632.

- Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, 5, 93-105.

et

Pour $\underline{w}_i = \underline{w}$ pour tout i et $\sigma_i^2 = \sigma^2 > 0$ pour tout i , nous obtenons

$$ieff(\mathbf{a}_i^*) = ieff(\mathbf{a}_i).$$

Pour $\underline{w}_i = \underline{w}$ pour tout i , nous obtenons

$$ieff(\mathbf{a}_i^*) \leq ieff(\mathbf{a}_i) \text{ ssi } \sum_i n_i \sigma_i^2 \sum_i n_i^2 \leq n \sum_i n_i^2 \sigma_i^2$$

Pour $\sigma_i^2 = \sigma^2 > 0$ pour tout i , nous obtenons

$$ieff(\mathbf{a}_i^*) \leq ieff(\mathbf{a}_i) \text{ ssi } n \sum_i n_i^2 \underline{w}_i^2 \leq \sum_i n_i \underline{w}_i^2 \sum_i n_i^2.$$

Résultat 4. Nous avons

$$ieff_w - ieff = \frac{SCT + n\bar{w}^2}{\bar{n}_{int}} \left[\sum_i \left(\frac{\bar{n}_{int}}{n_i} - 1 \right) n_i \underline{w}_i^2 - SCT \right] p_{int} = \frac{(1 + CV_{-2}^w) SCT}{\bar{n}_{int}} \left[\sum_i \left(\frac{\bar{n}_{int}}{n_i} - 1 \right) n_i \underline{w}_i^2 - SCT \right] p_{int} = \frac{1 + CV_{-2}^w}{\bar{n}_{int} \sigma_w^2} \left[\sum_i \left(\frac{\bar{n}_{int}}{n_i} - 1 \right) n_i \underline{w}_i^2 - SCT \right] p_{int}.$$

Preuve.

$$ieff_w - ieff$$

$$= \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int} = \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int} = \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int} = \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int}.$$

$$= \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int} = \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int} = \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int} = \frac{\left(\sum_i \left(\sum_k w_{ik} \right) \frac{1}{\bar{n}_{int}} - p_{int} \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)} p_{int}.$$

nous avons

$$\leq ieff(\mathbf{a}_i^*) \leq ieff(\mathbf{a}_i) \text{ si et uniquement si}$$

$$\sum_i n_i \sigma_i^2 \sum_i n_i^2 \underline{w}_i^2 \leq \sum_i n_i^2 \sigma_i^2 \sum_i n_i \underline{w}_i^2.$$

Preuve. Nous avons

$$ieff(\mathbf{a}_i^*) - ieff(\mathbf{a}_i) = \frac{\left(\sum_i n_i \sigma_i^2 \sum_i n_i^2 \underline{w}_i^2 - \sum_i n_i^2 \sigma_i^2 \sum_i n_i \underline{w}_i^2 \right)}{\left(\sum_i n_i^2 \underline{w}_i^2 - SCT + n\bar{w}^2 \right)}.$$

Pour $n_i = n/I$ pour tout i , nous obtenons

$$ieff(\mathbf{a}_i^*) = ieff(\mathbf{a}_i).$$

Pour $w_{ik} = \underline{w}_i$ pour tout i , c'est-à-dire $\sigma_i^2 = 0$, nous obtenons

$$ieff(\mathbf{a}_i^*) = ieff(\mathbf{a}_i).$$

et que

$$\text{Var}_{M_2}(\bar{y}^w) = \frac{\sigma^2 \left[\sum_{i=1}^I \sum_{k=1}^K w_{ik}^2 + \rho_{\text{int}} \sum_{i=1}^I \sum_{k \neq k'}^K w_{ik} w_{ik'} \right]}{\left(\sum_{i=1}^I \sum_{k=1}^K w_{ik}^2 \right)^2},$$

et en effectuant certaines opérations algébriques.

Corollaire : Supposons que $\rho_{\text{int}} > 0$ et $w_k = 1/n$. En utilisant le résultat 1 et l'inégalité de Cauchy-Schwarz, nous obtenons

$$\text{ieff}(\mathbf{a}_1) = 1 + \rho_{\text{int}} \left(\frac{\sum_{i=1}^I n_i^2}{n} - 1 \right) \geq 1 + \rho_{\text{int}} \left(\frac{I}{n} - 1 \right) = \text{ieff}.$$

Résultat 2. $\text{ieff}_w \leq \text{ieff}(\mathbf{a}_2)$, où

$$\mathbf{a}_2 = (a_{21}, \dots, a_{2I}) \text{ avec } a_{2i} = \frac{\sum_{k=1}^K w_{ik}^2}{\sum_{k=1}^K w_{ik}^2}.$$

Preuve : En utilisant l'inégalité de Cauchy-Schwarz, nous avons

$$\left(\sum_{k=1}^K w_{ik}^2 \right)^2 \leq \sum_{k=1}^I n_i \sum_{k=1}^K w_{ik}^2$$

avec égalité si et uniquement si $w_k = \bar{w}_i$ pour tout i et k , où

$$\bar{w}_i = \frac{\sum_{k=1}^K w_{ik}}{n_i}$$

est le poids de sondage moyen pour le $i^{\text{ème}}$ intervieweur. Donc, nous avons $\text{ieff}_w \leq 1 + [\bar{w}_{\text{int}}(\mathbf{a}_2) - 1] \rho_{\text{int}} = \text{ieff}(\mathbf{a}_2)$. L'égalité est vérifiée si et uniquement si $w_k = \bar{w}_i$ pour tout i et k , auquel cas $\text{ieff}_w = \text{ieff}(\mathbf{a}_2)$, où

$$\mathbf{a}_2 = (a_{21}, \dots, a_{2I}), \text{ avec } a_{2i} = \frac{\sum_{k=1}^I n_i w_{ik}^2}{n_i \bar{w}_i^2}.$$

Si tous les poids sont non négatifs, alors

$$\sigma_i^2 = \frac{1}{n_i} \sum_{k=1}^K (w_{ik} - \bar{w}_i)^2 \leq (n_i - 1) \bar{w}_i^2$$

puisque σ_i^2 est Schur-convexe. Définir

$$x_i = \frac{n_i}{1 + \frac{\bar{w}_i^2}{\sigma_i^2}} \text{ implique que } \frac{n_i}{1} \leq x_i \leq I$$

grande que la variabilité d'échantillonnage. Dans de

nombreuses enquêtes, ce genre d'évaluation, qui nécessite l'estimation des corrélations intra-intervieweur et intra-grappe, est difficile, voire même impossible, parce que les effets d'intervieweur sont souvent confondus avec les effets de grappes spatiales. L'utilisation d'un plan à échantillons supposés, proposé pour la première fois par Mahalanobis (1946), dans lequel les personnes interrogées sont affectées aléatoirement aux intervieweurs, est un moyen de contourner le problème. En pratique, la mise en œuvre d'un plan de ce genre dans une enquête par sondage à grande échelle est difficile, mais il est possible d'appliquer certains plans à échantillons supposés approximatifs (Hansen et coll. 1961, Bailar, Bailey et Stevens 1977, Bailey, Moore et Bailar 1978, Collins et Butcher 1982, O'Muircheartaigh et Campanelli 1998). Les modèles multinationaux ont été utilisés comme remède partiel au problème (Hox et De Leeuw 1994, Davis et Scott 1995, O'Muircheartaigh et Campanelli 1998, Scott et Davis 2001). Nous n'avons pas envisagé la question de l'estimation des corrélations intra-intervieweur et intra-grappe. Il s'agit d'un problème important dont nous traiterons dans un futur article.

En pratique, les effets d'intervieweur ou de plan sont calculés pour de nombreuses questions en utilisant la même formule, et une mesure sommaire, telle que l'effet médian d'intervieweur ou de plan, est utilisée pour la planification et la conception de l'enquête. En ce qui concerne les problèmes associés au traitement de questions multiples, le concepteur d'enquêtes peut poursuivre son propre protocole ; la seule modification que nous pourrions proposer consiste à utiliser nos nouvelles définitions de l'effet d'intervieweur ou de l'effet global dans la mesure du possible.

L'utilisation de notre formule peut suggérer des effets globaux qui sont nettement plus faibles que ceux indiqués par la formule standard. Cela, à son tour, pourrait suggérer une plus petite taille d'échantillon et, par conséquent, permettre de réaliser une économie.

Annexe

$$\text{Résultat 1. } \text{ieff}_w = \frac{\text{Var}_{M_2}(\bar{y}^w)}{\text{Var}_{M_1}(\bar{y}^w)} = 1 + \rho_{\text{int}} \left(\frac{\sum_{i=1}^I \left(\sum_{k=1}^K w_{ik}^2 \right)}{\left(\sum_{i=1}^I \sum_{k=1}^K w_{ik}^2 \right)^2} - 1 \right).$$

Preuve : Le résultat s'obtient en notant que

$$\text{Var}_{M_1}(\bar{y}^w) = \text{Var}_{M_1} \left[\frac{\sum_{i=1}^I \sum_{k=1}^K w_{ik} y_{ik}}{\sum_{i=1}^I \sum_{k=1}^K w_{ik}} \right] = \frac{\sigma^2 \sum_{i=1}^I \sum_{k=1}^K w_{ik}^2}{\left(\sum_{i=1}^I \sum_{k=1}^K w_{ik}^2 \right)^2},$$

Pour chacune des n_p personnes interrogées de l'UPE p et chacune des n_p' personnes interrogées de l'UPE p'

$$P(Y^{pik} = x_1, Y^{p'ik} = x_2)$$

$\begin{matrix} x_1 \\ \diagdown \\ x_2 \end{matrix}$	1	θ^2	θ	Total
	0	$\theta(1 - \theta)$	$1 - \theta$	$1 - \theta$
Total	θ	$\theta(1 - \theta)$	$1 - \theta$	1

Par conséquent, nous avons

$$E(Y^{pik}) = \theta \text{ pour tout } p, i, k,$$

$$\text{Var}(Y^{pik}) = \theta(1 - \theta) \text{ pour tout } p, i, k,$$

$$\rho = \frac{\text{Cov}(Y^{pik}, Y^{p'ik})}{\sqrt{\text{Var}(Y^{pik})\text{Var}(Y^{p'ik})}}$$

$$= \frac{\alpha - \theta^2}{\theta} \text{ pour tout } p, i \text{ et } k \neq k',$$

$$\rho_C = \frac{\text{Cov}(Y^{pik}, Y^{p'ik'})}{\sqrt{\text{Var}(Y^{pik})\text{Var}(Y^{p'ik'})}}$$

$$= \frac{\beta - \theta^2}{\theta(1 - \theta)} \text{ pour tout } p \text{ et } i \neq i',$$

ce qui est un cas particulier du modèle M_4 avec $\sigma^2 = \text{Var}(Y^{pik}) = \theta(1 - \theta)$. Notons que ρ_C ainsi que ρ peuvent être négatifs et que $\rho_{\text{int}} = \rho - \rho_C$ est positif si et uniquement si $\alpha > \beta$.

Remarque 5.5 : Pour un plan EPE avec taille d'UPE commune $b = n/P$, nous avons

$$eff = 1 + \rho_C(b - 1) + \rho_{\text{int}}(A - 1).$$

Remarque 5.6 : Dans sa discussion de Verma et coll. (1980), Holt a considéré le cas où il n'existe aucune variabilité d'intervieweur et où l'UPE est la classe de pondération, c'est-à-dire le cas où $\rho_{\text{int}} = 0$ et $w_{pik} = w_p$ pour tout p, i, k . Dans ces conditions, eff se réduit à

$$eff = \frac{\left(\sum_{d=1}^D n^d w^d\right) \left(\sum_{d=1}^D n^d w^d\right)}{2} \times \left[1 + \rho_C \frac{\left(\sum_{d=1}^D n^d w^d\right) \left(\sum_{d=1}^D n^d w^d\right)}{\left(\sum_{d=1}^D n^d w^d\right) \left(\sum_{d=1}^D n^d w^d\right)} - 1 \right].$$

effectuant certaines opérations algébriques. Les formules de l'effet de plan en l'absence d'effets d'intervieweur ont été examinées par de nombreux auteurs. Voir Kish (1965), Verma et coll. (1980), Skinner (1986), Valliant (1987), Skinner et coll. (1989), Gabler, Häder et Lahiri (1999), Lynn et Gabler (2004), Kalton, Brick et Lé (2005), et d'autres.

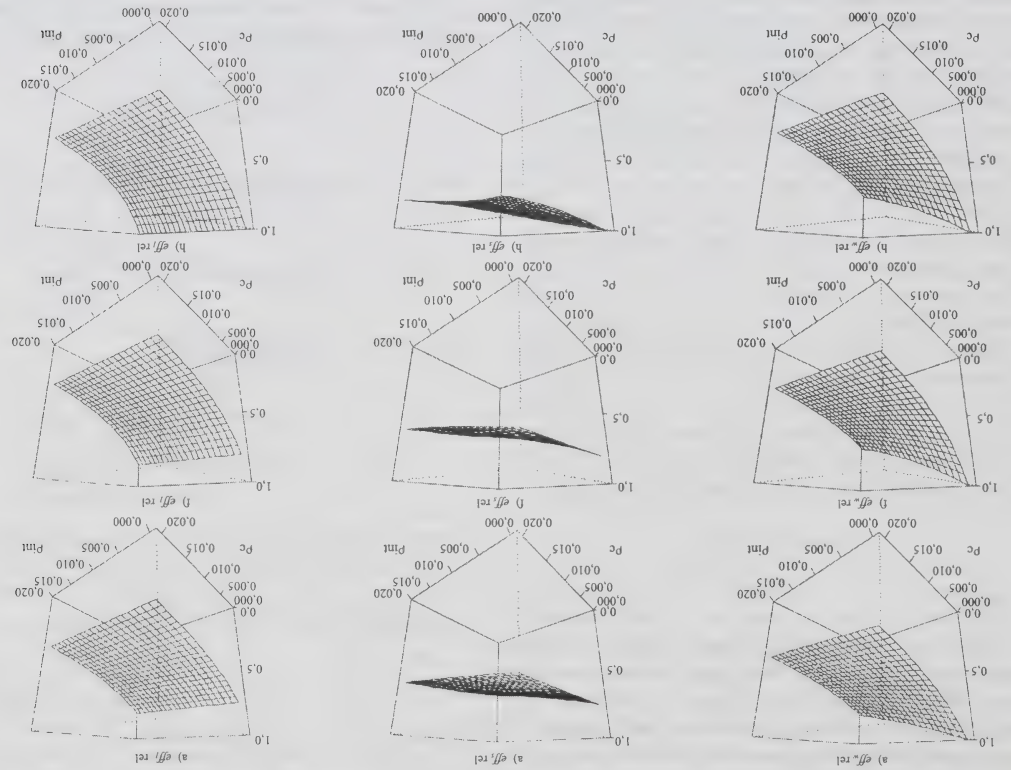
6. Conclusion

Nous avons constaté que la formule standard de l'effet d'intervieweur pourrait poser un problème de surestimation ou de sous-estimation selon la situation. Par exemple, elle pourrait sous-estimer gravement les effets d'intervieweur dans le cas d'un plan d'échantillonnage avec probabilité égale (EPE) quand les charges de travail des intervieweurs sont différentes. Curieusement, la corrélation spatiale peut transformer ce sous-estimation en une surestimation. Dans le premier cas, le concepteur d'enquêtes qui utilise la formule standard d'effet d'intervieweur pourrait ne pas accorder suffisamment d'attention au contrôle de cet effet. Dans le second cas, une valeur élevée de l'effet d'intervieweur pourrait susciter inutilement des préoccupations au sujet de la qualité des données associées à l'intervieweur. Cela pourrait déclencher l'affectation d'une part plus grande qu'il n'est nécessaire du budget à la réduction de l'effet d'intervieweur, lequel pourrait déjà être nettement plus faible que la valeur obtenue par une application de la formule standard. L'objectif de l'article est de définir et d'interpréter les effets d'intervieweur qui sont appropriés dans diverses situations d'enquêtes complexes.

Nous avons considéré le cas où un intervieweur est affecté à une seule UPE. Le cas où un intervieweur travaille dans différentes UPE est également important et sera examiné dans un futur article. Les poids utilisés dans les formules proposées tiennent compte uniquement des poids d'échantillonnage, car ils sont plantifiés à l'étape de la conception de l'enquête, mais ne reflètent pas nécessairement les poids réels associés à chaque cas, une fois que les données sont recueillies. Autrement dit, nos formules d'effet d'intervieweur ne tiennent pas compte des effets dus à la non-réponse et aux corrections par poststratification. Les formules présentées dans l'article sont surtout utiles à l'étape de la planification et de la conception, quand nous avons une certaine idée des corrélations intra-intervieweur et spatiale.

L'estimation fiable de ρ_{int} et ρ_C est importante. Bien que certains articles publiés traitent de cette estimation, il est sans aucun doute nécessaire de faire progresser la recherche dans ce domaine important. En comparant les deux sources d'homogénéité, Hansen et coll. (1961) ont constaté que la variabilité d'intervieweur était souvent plus

Figure 2 Contributions relatives des effets de pondération, de plan et d'intervieweur à l'effet global pour les cas a), f) et h)



Pour le modèle linéaire mixte susmentionné, il est facile de vérifier que

$$P_{mi} = \frac{\sigma^2_{\alpha}}{\sigma^2_{\alpha} + \sigma^2_{\beta} + \sigma^2_{\epsilon}} \text{ et } P_c = \frac{\sigma^2_{\alpha}}{\sigma^2_{\alpha} + \sigma^2_{\beta} + \sigma^2_{\epsilon}}.$$

Cependant, il est intéressant de souligner que la définition ne nécessite pas que P_{mi} et P_c soient strictement positifs et elle s'applique à l'exemple suivant :

Exemple 3 : Modèle simple pour des données binaires.

En supposant que $0 < \min(\alpha, \beta) < \theta < 1$, nous définissons le modèle suivant :

Pour chacune des n_{pi} personnes interrogées par l'intervieweur i dans l'UPE p ,

x_1 \ x_2	1	β	θ	Total
	0	$\theta - \alpha$	$1 - 2\theta + \beta$	$1 - \theta$
1	θ	$\theta - \alpha$	$1 - \theta$	1
Total	0	1	$1 - \theta$	1

$$P(Y_{piik} = x_1, Y_{piik'} = x_2)$$

Pour chacune des n_{pi} personnes interrogées par l'intervieweur i dans l'UPE p et chacune des $n_{pi'}$ personnes interrogées par l'intervieweur i' dans l'UPE p ,

x_1 \ x_2	1	α	θ	Total
	0	$\theta - \alpha$	$1 - 2\theta + \alpha$	$1 - \theta$
1	θ	$\theta - \alpha$	$1 - \theta$	1
Total	0	1	$1 - \theta$	1

$$P(Y_{piik} = x_1, Y_{pi'k'} = x_2)$$

Tableau 3
Charges de travail d'intervieweur moyennes pour plusieurs combinaisons de paramètres (exemple 2) : $ieff_w / ieff_{s,w}$ pour $p_{int} = 0,01$ et $p_c = 0,02$

$IA = (3,1)$												
n_i	w_i	σ_i^2	$\bar{n}_{int}(A_*)$	$\bar{n}_{sp}(B_*)$	p_c	$\frac{ieff_{s,w}}{ieff_{s,n}}$	$\bar{n}_{int}(A_*)$	$\bar{n}_{sp}(B_*)$	p_c	$\frac{ieff_{s,w}}{ieff_{s,n}}$	$\bar{n}_{int}(A_*)$	$\bar{n}_{sp}(B_*)$
a)	25	1,022	0,299	25	19,202	47,528	-0,005	1,133	19,202	38,389	-0,006	1,123
	25	0,945	0,260									
	25	1,036	0,375									
b)	10	1	1	25	15	41	-0,010	1,151	15	29	-0,015	1,138
	20	1	1									
	30	1	1									
	40	1	1									
c)	10	1	1	25	7,5	20,5	-0,037	1,185	7,5	14,5	-0,054	1,180
	20	1	1									
	30	1	1									
	40	1	1									
d)	10	1	1	25	10	27,333	-0,024	1,171	10	19,333	-0,034	1,163
	20	1	1									
	30	1	1									
	40	1	1									
e)	10	4	144	25	1,801	2,755	-0,551	1,230	1,801	3,603	-0,371	1,231
	20	2	9									
	30	0,333	0,555									
	40	0,250	0,125									
f)	10	0,333	0,025	25	31,820	75,685	0,004	1,104	31,820	58,427	0,005	1,084
	20	0,666	0,075									
	30	0,125	0,125									
	40	1,333	0,175									
g)	10	1	0,010	25	29,126	79,612	0,002	1,118	29,126	56,311	0,003	1,094
	20	0,020	0,020									
	30	1	0,030									
	40	1	0,040									
h)	10	1	0,004	25	29,940	81,836	0,003	1,117	29,940	57,884	0,004	1,084
	20	1	0,003									
	30	1	0,002									
	40	1	0,001									

5. Effet global

L'effet global tient compte de la pondération inégale, des grappes spatiales et des effets d'intervieweur, et peut être considéré comme une généralisation de l'effet de plan classique. La multiplication de la variance sous EAS pour la moyenne d'échantillon non pondérée par l'effet global fournira l'estimateur de variance totale

$$eff = \frac{Var_{M_i}(\bar{y}_w)}{Var_{M_i}(\bar{y})} = eff_w \times eff_s \times eff_{int},$$

où

$$eff_w = \frac{Var_{M_i}(\bar{y}_w)}{Var_{M_i}(\bar{y})}, \quad eff_s = \frac{Var_{M_i}(\bar{y}_w)}{Var_{M_i}(\bar{y}_w)}, \quad eff_{int} = \frac{Var_{M_i}(\bar{y}_w)}{Var_{M_i}(\bar{y}_w)}.$$

Les contributions relatives de la pondération, des grappes spatiales et des effets d'intervieweur à l'effet global sont données par

$$eff = \frac{n \sum_p \sum_l \sum_{i=1}^p \sum_{k=1}^n w_{pik}^2}{\sum_{p=1}^p \sum_{l=1}^l \sum_{i=1}^p \sum_{k=1}^n w_{pik}^2} \times \left[1 + p_c \left(\frac{\sum_{p=1}^p \sum_{l=1}^l \sum_{i=1}^p \sum_{k=1}^n w_{pik}^2}{\sum_{p=1}^p \sum_{l=1}^l \sum_{i=1}^p \sum_{k=1}^n w_{pik}^2} \right) - 1 \right] + p_{int} \left(\frac{\sum_{p=1}^p \sum_{l=1}^l \sum_{i=1}^p \sum_{k=1}^n w_{pik}^2}{\sum_{p=1}^p \sum_{l=1}^l \sum_{i=1}^p \sum_{k=1}^n w_{pik}^2} \right) - 1$$

Nous pouvons montrer que

$$M_i^* : Cov(y_{pik}, y_{p'ik'}) = \begin{cases} \sigma^2 & \text{si } p = p', i = i', k = k', \\ 0 & \text{autrement.} \end{cases}$$

suivant :

Ci-dessus, $Var_{M_i^*}$ est calculé par rapport au modèle

Remarque 4.2 : Définissons

$$\bar{n}_{int}^{(A)} = \sum_{l_p}^p \sum_{l_j}^p a^{l_p l_j} n^{p_{lj}}, \text{ où } A = ((a^{l_p l_j})), \text{ avec } a^{l_p l_j} = \frac{n}{n^{p_{lj}}},$$

et

$$\bar{n}_{upc}^{(b)} = \sum_p^p b^{n_p}, \text{ ou } b = (b^1, \dots, b^p) \text{ avec } b^p = \frac{n}{n^p}.$$

Si $p_c \neq 0$, mais que nous utilisons un plan EPE, nous laissons tomber l'indice inférieur w dans $ieff_{s,w}$. Notons

que

$$ieff_s = 1 + p_{int} \frac{\bar{n}_{int}^{(A)} - 1}{1 + p_{int} [\bar{n}_{upc}^{(b)} - 1]}$$

de sorte que

$$= 1 + p_{int} \frac{p_c}{1 + p_{int}} \cdot \frac{\bar{n}_{int}^{(A)} - 1}{p_c} \cdot \frac{\bar{n}_{upc}^{(b)} - 1}{1 + p_c [\bar{n}_{upc}^{(b)} - 1]}$$

$$ieff_s > 1 + p_{int} \frac{p_c}{1 + p_{int}} \cdot \frac{\bar{n}_{int}^{(A)} - 1}{p_c} \cdot \frac{\bar{n}_{upc}^{(b)} - 1}{1 + p_c} \cdot \frac{\bar{n}_{int}^{(A)}}{\bar{n}_{upc}^{(b)}}.$$

Il est facile de voir que le deuxième membre de l'inégalité

$$\frac{\bar{n}_{int}^{(A)}}{\bar{n}_{upc}^{(b)}} - 1$$

augmente avec les ratios p_{int}/p_c et

Nous avons

$$\frac{\bar{n}_{int}^{(A)} - 1}{1 + p_c} - [1 + p_c (\bar{n}_{upc}^{(b)} - 1)] \frac{ieff_s - ieff}{\bar{n}_{int}^{(A)} - 1} = p_{int} \frac{1 + p_c (\bar{n}_{upc}^{(b)} - 1)}{(\bar{n}_{int} - 1)}.$$

Donc, pour $p_{int} > 0$,

$ieff_s < ieff$ si et uniquement si

$$Deff_s := 1 + p_c (\bar{n}_{upc}^{(b)} - 1) > \frac{\bar{n}_{int}^{(A)} - 1}{\bar{n}_{int} - 1},$$

c'est-à-dire si et uniquement si l'effet de plan dû aux grappes spatiales est plus grand que le ratio de la moyenne pondérée des charges de travail d'intervieweur -1 à la charge de travail est la même pour tous les intervieweurs, le deuxième membre de l'inégalité est égal à 1 et, par conséquent, l'inégalité est toujours valide. Il est intéressant de souligner que $ieff \approx 4 \cdot ieff_s$ si $p_{int} = 0,1$, $p_c = 0,05$, $\bar{n}_{upc}^{(b)} = 140$, et $\bar{n}_{int} = 70$.

Remarque 4.3 : Dans le cas général, nous avons

$$ieff_{s,w} - ieff = p_{int} \left(\frac{\bar{n}_{int}^{(A^w)} - 1}{1 + p_c (\bar{n}_{upc}^{(b^w)} - 1)} - (\bar{n}_{int} - 1) \right).$$

Donc, pour $p_{int} > 0$,

$ieff_{s,w} < ieff$ si et uniquement si

$$Deff_{s,w} := 1 + p_c (\bar{n}_{upc}^{(b^w)} - 1) > \frac{\bar{n}_{int}^{(A^w)} - 1}{\bar{n}_{int} - 1},$$

$$p_c > \frac{\bar{n}_{int}^{(A^w)} - 1}{\bar{n}_{int} - 1} (\bar{n}_{upc}^{(b^w)} - 1) =: p_c^*, \text{ disons.}$$

Dans l'exemple 2 (voir le tableau 3), $ieff$ est une valeur prudente de $ieff_{s,w}$ (pour a) à e) si $p_c > 0$. Il en est de même pour f) à h) si $p_c > 0,004$.

Si un ménage et une personne dans ce ménage sont sélectionnés au hasard, les poids sont souvent indépendants de l'UPB et de l'intervieweur, et dépendent uniquement de la taille du ménage. Le cas échéant, les tailles de ménage forment les classes de pondération. Pour ces dernières, nous définissons

m_{pj}^p : nombre d'unités d'échantillonnage dans l'UPB p affectées à l'intervieweur j appartenant à la classe de pondération j .

$m_{pj}^p = \sum_{i=1}^p m_{pji}^p$: nombre d'unités d'échantillonnage dans l'UPB p appartenant à la classe de pondération j .

$m_j = \sum_{p=1}^p \sum_{i=1}^p m_{pji}^p$: nombre d'unités d'échantillonnage appartenant à la classe de pondération j .

Donc,

$n_{pj}^p = \sum_{i=1}^p m_{pji}^p$: nombre d'unités d'échantillonnage dans l'UPB p affectées à l'intervieweur i ,
 $n_p^p = \sum_{i=1}^p \sum_{j=1}^p m_{pji}^p$: nombre d'unités d'échantillonnage dans l'UPB p ,
 $n = \sum_{p=1}^p \sum_{i=1}^p \sum_{j=1}^p m_{pji}^p$: taille de l'échantillon.

En outre,

$$\bar{n}_{int}^{(A^w)} = \frac{\sum_{l_p}^p \sum_{l_j}^p \sum_{k=1}^p w^{p_{lk}}}{\sum_{l_p}^p \sum_{l_j}^p \sum_{k=1}^p w^{p_{lk}}} = \frac{\sum_{l_p}^p \sum_{l_j}^p \sum_{k=1}^p w^{p_{lk}}}{\sum_{l_p}^p \sum_{l_j}^p \sum_{k=1}^p w^{p_{lk}}} = \frac{\sum_{l_p}^p \sum_{l_j}^p \sum_{k=1}^p w^{p_{lk}}}{\sum_{l_p}^p \sum_{l_j}^p \sum_{k=1}^p w^{p_{lk}}}$$

et

sont les ratios des formes quadratiques dans $w = (w_1, \dots, w_j)$.

4. Pondération inégale et grappes spatiales

À la présente section, nous obtenons une formule appropriée de la variance d'intervaleur en présence de grappes spatiales et de probabilités inégales de sélection. Considérons la situation où plus d'un intervieweur travaillent indépendamment dans la même UPE et où, dans chaque UPE, les enquêtes sont affectées aléatoirement aux intervieweurs. Nous supposons qu'aucun intervieweur ne travaille dans plus d'une UPE. Une telle situation a été considérée par Biemer et Stokes (1985). Maintenant, nous allons séparer l'effet d'intervaleur de l'effet d'UPE (c'est-à-dire l'existence de grappes spatiales) et de la pondération inégale. Soit y_{pik} et w_{pik} l'observation et le poids de sondage connexe pour la $k^{\text{ième}}$ personne dans la $p^{\text{ième}}$ UPE interrogée par le $i^{\text{ième}}$ intervieweur ($p = 1, \dots, P$; $i = 1, \dots, I_p$; $k = 1, \dots, n_{pi}$). Soit $n_p = \sum_{i=1}^{I_p} n_{pi}$ le nombre d'unités d'échantillonnage dans l'UPE p .

Dans ce cas, nous utilisons la moyenne pondérée qui suit pour estimer la moyenne de population finie :

$$\bar{y}_w = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} y_{pik}}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}}.$$

Définitions

$$eff_{s,w} = \frac{Var_{M_j}(\bar{y}_w)}{Var_{M_j}(\bar{y}_w)}.$$

où les indices inférieurs s et w signifient la présence de grappes spatiales et de poids inégaux. Dans l'expression ci-dessus, $Var_{M_j}(\bar{y}_w)$ et $Var_{M_j}(\bar{y}_w)$ sont les variances de \bar{y}_w sous les deux modèles suivants, respectivement

$$M_3 : Cov(y_{pik}, y_{p'ik'}) = \begin{cases} \sigma^2 & \text{si } p = p', i = i', k = k' \\ p \cdot \sigma^2 & \text{si } p = p', k \neq k' \\ 0 & \text{autrement} \end{cases}$$

$$M_4 : Cov(y_{pik}, y_{p'ik'}) = \begin{cases} \sigma^2 & \text{si } p = p', i = i', k = k' \\ p \cdot \sigma^2 & \text{si } p = p', i \neq i' \\ p \sigma^2 & \text{si } p = p', i = i', k \neq k' \\ 0 & \text{si } p \neq p' \end{cases}$$

Dans ces modèles, p_C est la corrélation intra-UPE et p_{int} est la corrélation intra-classe combinée d'intervaleurs et d'UPE. Définissons $p_{int} = p - p_C$, la corrélation intra-intervaleur. Habituellement, $p_{int} > 0$.

Du résultat 5, il découle que

$$eff_{s,w} = 1 + p_{int} \frac{\bar{n}_{int}(\mathbf{A}_w) - 1}{1 + p_C(\bar{n}_{upc}(\mathbf{b}_w) - 1)},$$

$$\mathbf{A}_w = ((a_{wpil})_{i=1, \dots, I_p}^{p=1, \dots, P}) \text{ et } a_{wpil} = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{n_{pi} w_{pi}^2}.$$

$$\bar{w}_{pi} = \frac{1}{n_{pi}} \sum_{k=1}^{n_{pi}} w_{pik}$$

$$\bar{n}_{int}(\mathbf{A}_w) = \sum_{p=1}^P \sum_{i=1}^{I_p} a_{wpil} n_{pi} = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}.$$

$$\mathbf{b}_w = (b_{wpil})_{p=1, \dots, P} \text{ et } b_{wpil} = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{n_{pi} w_{pi}^2},$$

$$\bar{w}_p = \frac{1}{I_p} \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik} = \frac{1}{I_p} \sum_{i=1}^{I_p} n_{pi} \bar{w}_{pi}$$

$$\bar{n}_{upc}(\mathbf{b}_w) = \sum_{p=1}^P b_{wpil} n_p = \frac{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}{\sum_{p=1}^P \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2}.$$

Notons que $\bar{n}_{int}(\mathbf{A}_w) \leq \bar{n}_{upc}(\mathbf{b}_w)$ avec égalité si et uniquement si $I_p = 1$. Notons aussi que $\bar{n}_{int}(\mathbf{A}_w)$ ne varie pas en fonction de l'affectation des intervieweurs aux UPE, tandis que $\bar{n}_{upc}(\mathbf{b}_w)$ varie.

Remarque 4.1 : Si $p_C = 0$, nous obtenons

$$eff_{s,w} = 1 + p_{int}(\bar{n}_{int}(\mathbf{A}_w) - 1).$$

Cette formule est semblable à eff_w donné à la section 2. Donc, tous les commentaires formulés à la remarque 2.1 s'appliquent ici. Notons que $\bar{n}_{int}(\mathbf{A}_w)$, tout comme $\bar{n}_{upc}(\mathbf{A}_w)$, ne peut généralement pas s'écrire sous la forme $\bar{n}_{int}(\mathbf{A}_w) = \sum_{p=1}^P \sum_{i=1}^{I_p} a_{wpil} n_{pi}$ avec $\sum_{p=1}^P \sum_{i=1}^{I_p} a_{wpil} = 1$; le même commentaire s'applique à $\bar{n}_{upc}(\mathbf{b}_w)$.

- c) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. En outre, puisque $\bar{w}_i^2 + \sigma_i^2$ et n_i augmentent tous deux, nous avons $ieff(a_1) < ieff(a_2)$.
- d) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. Puisque $\bar{w}_i^2 + \sigma_i^2$ diminue et que n_i augmente, nous avons $ieff(a_2) < ieff(a_1)$.
- e) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. En outre, \bar{w}_i^2 et σ_i^2 diminuent et n_i augmente, ce qui implique que $ieff(a_2) < ieff(a_1)$.

- f) Le fait que \bar{w}_i^2 et n_i augmentent implique que $ieff_w > ieff$; puisque σ_i^2 et n_i augmentent tous deux, nous avons $ieff(a_1) < ieff(a_2)$.
- g) Puisque \bar{w}_i^2 et n_i augmentent, nous avons $ieff_w > ieff$ et, puisque σ_i^2 augmente, nous avons $ieff(a_1) < ieff(a_2)$. En outre, $ieff_w < ieff(a_1)$ puisque σ_i^2 est plus petit que dans f).
- h) Puisque \bar{w}_i^2 et n_i augmentent, nous avons $ieff_w > ieff$ et puisque σ_i^2 diminue, nous avons $ieff(a_2) < ieff(a_1)$.

Tableau 2
Classement des formules d'effet d'intervieweur pour plusieurs combinaisons de paramètres
(exemple 2) : dans la dernière colonne $p_{int} = 0,01$

	n_i	\bar{w}_i^2	σ_i^2	$\bar{n}_{int}(a_1)$	$\bar{n}_{int}(a_2)$	\bar{n}_w	Effets d'intervieweur	$ieff / ieff_w$
a)	25	1,022	0,299	25	25	25	$ieff = ieff(a_1) = ieff(a_2) > ieff_w$	1,003
b)	10	1	1	25	30	15	$ieff_w < ieff < ieff(a_1) = ieff(a_2)$	1,007
	20	1	1					
	30	1	1					
	40	1	1					
c)	10	1	1	25	30	7,5	$ieff_w < ieff < ieff(a_1) < ieff(a_2)$	1,023
	20	1	2					
	30	1	3					
	40	1	4					
d)	10	1	4	25	30	10	$ieff_w < ieff < ieff(a_1) < ieff(a_2)$	1,015
	20	1	3					
	30	1	2					
	40	1	1					
e)	10	4	144	25	30	1,80	$ieff_w < ieff < ieff(a_2) < ieff(a_1)$	0,998
	20	2	9					
	30	0,333	0,555					
	40	0,250	0,125					
f)	10	0,333	0,025	25	30	31,82	$ieff < ieff_w < ieff(a_1) < ieff(a_2)$	1,015
	20	0,666	0,075					
	30	1	0,125					
	40	1,333	0,175					
g)	10	1	0,010	25	30	29,13	$ieff < ieff_w < ieff(a_1) < ieff(a_2)$	0,999
	20	1	0,020					
	30	1	0,030					
	40	1	0,040					
h)	10	1	0,004	25	30	29,94	$ieff < ieff_w < ieff(a_2) < ieff(a_1)$	0,998
	20	1	0,003					
	30	1	0,002					
	40	1	0,001					

$SCE = \sum_{i=1}^I n_i (\bar{w}_i - \bar{w})^2$, la somme des carrés entre intervieweurs des poids de sondage,

$SCI = \sum_{i=1}^I n_i \sum_{j=1}^{n_i} (w_{ij} - \bar{w}_i)^2 = \sum_{i=1}^I n_i \sigma_i^2$, la somme des carrés intra-intervieweur des poids de sondage,

$SCt = SCE + SCI$, la somme totale des carrés des poids de sondage,

$\tau_w = SCI/SCt$, un indicateur de la contribution relative de la variabilité intra-intervieweur des poids de sondage à la variabilité totale,

$CV_w = \sqrt{SCt/n} / \bar{w}$, le coefficient de variation des poids de sondage dans l'échantillon complet.

Nous pouvons montrer que (voir le résultat 4)

$$ieff_w - ieff$$

$$(1) \quad = \frac{\bar{w}_{int}^{SCt + m\bar{w}_i^2}}{\sum_{i=1}^I \left(\frac{n_i}{n_i \bar{w}_i^2} - 1 \right) n_i \bar{w}_i^2 - SCI} \left[p_{int} \right]$$

$$(2) \quad = \frac{\bar{w}_{int}^{(1 + CV_w^{-2}) SCt}}{\sum_{i=1}^I \left(\frac{n_i}{n_i \bar{w}_i^2} - 1 \right) n_i \bar{w}_i^2 - SCI} \left[p_{int} \right]$$

$$(3) \quad = \frac{1 + CV_w^{-2}}{\bar{w}_{int} \tau_w} \left(\sum_{i=1}^I \left(\frac{n_i}{n_i \bar{w}_i^2} - 1 \right) \frac{SCt}{n_i \bar{w}_i^2} - 1 \right) \left[p_{int} \right]$$

Remarque 3.2 : Nous pouvons utiliser la formule (1) dans toute situation. Pour les plans EPE, nous avons

$$ieff_w - ieff = p_{int} \frac{n}{\sum_{i=1}^I \left(\frac{n_i}{n_i \bar{w}_i^2} - 1 \right) n_i} \left(\frac{n_{int}}{n_i \bar{w}_i^2} - 1 \right) n_i.$$

Notons qu'une application de l'inégalité de Cauchy-Schwarz donne à penser que $ieff_w - ieff \geq 0$ avec égalité si et uniquement si $n_i = n/I$ pour tout i .

Remarque 3.3 : Nous pouvons utiliser (2) si $SCt \neq 0$, c'est-à-dire si le plan n'est pas EPE. Si $p_{int} > 0$, implique que

$$ieff_w - ieff \leq 0 \text{ si et uniquement si } \sum_{i=1}^I \left(\frac{n_i}{n_i \bar{w}_i^2} - 1 \right) n_i \bar{w}_i^2 \leq SCt.$$

Si une charge d'intervieweur élevée a tendance à être associée à de petits poids moyens de sondage et inversement et que $SCt \neq 0$, nous pouvons nous attendre à ce que $ieff$ soit une valeur prudente de la variance d'intervieweur réelle

$ieff_w$. Dans l'exemple 2, c) et d), nous sommes en présence d'une telle situation.

Malheureusement, nous avons $ieff_w = ieff$ si et uniquement si $w_{ik} = \bar{w}_i$ (ou, de manière équivalente, $SCt = 0$) et $n_i \bar{w}_i^2 / \sum_{i=1}^I n_i \bar{w}_i^2 = 1/I$ pour tout i et k , c'est-à-dire $ieff_w = ieff$, si et uniquement si $w_{ik} = \bar{w}_i \propto 1/\sqrt{n_i}$ pour tout i et k .

Donc, pour un plan non EPE, une charge de travail d'intervieweur égale ne fournit pas nécessairement une interprétation assistée par modèle de $ieff$. Par exemple, si les poids de sondage varient pour au moins un intervieweur, nous n'aurons pas d'interprétation assistée par modèle de $ieff$. Évidemment, pour un plan EPE, les deux formules sont équivalentes si et uniquement si les charges de travail des intervieweurs sont égales.

Remarque 3.4 : Si la charge de travail est la même pour tous les intervieweurs, nous avons

$$ieff_w - ieff = - \frac{1 + CV_w^{-2}}{\bar{w}_{int} \tau_w} p_{int}$$

(en supposant que $SCt \neq 0$). Donc, $ieff$ est une valeur prudente de l'effet réel d'intervieweur $ieff_w$. En outre, $|ieff_w - ieff|$ est une fonction croissante de la charge de travail d'intervieweur commune \bar{w}_{int} et $\tau_w / (1 + CV_w^{-2})$ (pour CV_w^{-2} constant, cette dernière expression est une fonction croissante de τ_w). La même charge de travail d'intervieweur est donnée dans l'exemple 2 a).

Remarque 3.5 : Nous pouvons utiliser la formule (3) si $SCt > 0$, c'est-à-dire s'il existe au moins un intervieweur pour lequel les poids ne sont pas égaux.

Exemple 2.

Le tableau 2 donne huit combinaisons différentes de $(n_i, \bar{w}_i, \sigma_i^2)$. La première suppose que les valeurs de n_i sont égales, mais que les poids sont inégaux. La deuxième suppose que $\bar{w}_i^2 \propto \sigma_i^2$. Les six autres combinaisons montrent tous les classements possibles de \bar{w}_{int} , $ieff_w$, $n_i \bar{w}_i^2 / \sum_{i=1}^I n_i \bar{w}_i^2$ et, par conséquent $ieff(a_1)$, $ieff_w$, $ieff(a_2)$ si l'on tient compte du fait que $ieff \leq ieff(a_1)$ et $ieff_w \leq ieff(a_2)$.

Dans l'exemple, $\sum_{i=1}^I n_i \bar{w}_i = n$. Nous expliquons maintenant les huit profils différents.

- Puisque tous les n_i sont égaux, $ieff = ieff(a_1) = ieff(a_2)$. En outre, $ieff_w$ est plus petit que tous les autres, parce que $\sigma_i^2 > 0$.
- Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff(a_1) = ieff(a_2)$. En outre, $\sigma_i^2 = c \cdot \bar{w}_i^2$ implique que $ieff(a_1) = ieff(a_2)$.

valeurs de la variance d'intervieweur obtenues au moyen de la formule standard (c'est-à-dire, $ieff_w$) et de notre formule de variance d'intervieweur assistée par modèle pour toutes les combinaisons des deux facteurs influents, c'est-à-dire la moyenne pondérée des charges de travail d'intervieweur et la corrélation intra-intervieweur. En examinant la figure 1, il est intéressant de noter que $ieff_w$ pourrait donner une sous-estimation d'environ 100 %.

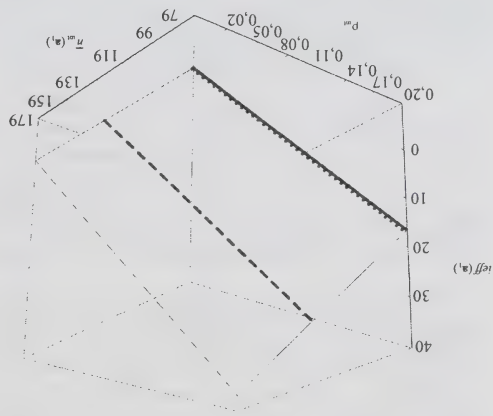


Figure 1 Graphiques de $ieff(a_i)$ en fonction de p_{int} pour différents $\bar{\pi}_{int}(a_i)$
Intervieweur A : trait interrompu
Intervieweur B : pointillé
Intervieweur C : trait plein

3. Pondération inégale sans grappes spatiales

À la présente section, nous examinons la situation où les poids sont inégaux. Soit w_k le poids de sondage associé à la $k^{ième}$ personne interrogée par le $i^{ième}$ intervieweur. Dans cette situation, la moyenne pondérée $\bar{y}_w = \sum_i \sum_k w_k y_{ik} / \sum_i \sum_k w_k$ est un estimateur fréquemment utilisé de la moyenne de population finie (voir Brewer 1963 ; Hájek 1971) et la formule de la variance d'intervieweur assistée par modèle est donnée par

$$ieff_w = \frac{\text{Var}_{M_i}(\bar{y}_w)}{\text{Var}_{M_i}(\bar{y})} = 1 + p_{int} \left(\frac{\sum_i \left(\sum_k w_k \right)^2}{\sum_i \sum_k w_k^2} - 1 \right).$$

Voit le résultat 1 en annexe.

Définissons $\bar{w}_i = 1/n_i \cdot \sum_{k=1}^{n_i} w_k$, le poids de sondage moyen pour le $i^{ième}$ intervieweur et $\sigma_i^2 = 1/n_i \cdot \sum_k w_k^2 - \bar{w}_i^2$, la variance des poids de sondage pour le $i^{ième}$ intervieweur.

On peut montrer que

$$ieff_w = 1 + p_{int}(\bar{w}_w - 1),$$

où

$$\bar{w}_w = \frac{\sum_i n_i \bar{w}_i^2}{\sum_i n_i \sigma_i^2 + \sum_i n_i \bar{w}_i^2}.$$

Notons qu'en général, $ieff_w$ ne peut pas s'écrire sous la forme $ieff_w = 1 + p_{int}(\bar{\pi}_{int}(a) - 1)$ avec $\sum_i a_i = 1$.

Remarque 3.1 : Du résultat 2 en annexe, il découle que

$$ieff_w \leq ieff(a_i),$$

où

$$a_i = (a_{21}, \dots, a_{2I}), \text{ avec } a_{2i} = \frac{\sum_k w_k^2}{\sum_i \sum_k w_k^2}.$$

Dans l'équation qui précède, pour $p_{int} > 0$, $ieff_w = ieff(a_i)$ si et uniquement si tous les σ_i^2 sont nuls. Donc, $ieff(a_i)$ peut être interprété comme une variance d'intervieweur prudente.

L'égalité est vérifiée si et uniquement si $w_k = \bar{w}_i$ pour tout i et k , auquel cas

$$ieff_w = ieff(a_i),$$

où

$$a_i = (a_{21}, \dots, a_{2I}), \text{ avec } a_{2i} = \frac{\sum_i n_i \bar{w}_i^2}{n_i \bar{w}_i^2}.$$

Donc, les formules $ieff_w$ et $ieff(a_i)$ sont équivalentes si et uniquement si les poids de sondage sont tous les mêmes pour un intervieweur donné. Un exemple de plan de sondage de ce genre est un plan EPF pour lequel nous avons

$$a_{2i} = \frac{n}{n_i}$$

et

$$ieff_w = ieff(a_i) = ieff(a).$$

Maintenant, nous allons essayer de voir quels facteurs expliquent la différence entre $ieff_w$ et $ieff$. Pour cela, définissons

$\bar{w} = 1/n \cdot \sum_{i=1}^I \sum_{k=1}^{n_i} w_k = \sum_{i=1}^I n_i / n \cdot \bar{w}_i$, le poids de sondage moyen pour l'ensemble des intervieweurs,

spatiales. Dans le cas de grappes spatiales, nous soutenons que la formule $ieff$ a généralement tendance à surestimer la variabilité d'intervieweur. Donc, pour les enquêtes complexes avec grappes spatiales, $ieff$ peut donner inutilement une fausse alarme quant à l'importance de la variabilité d'intervieweur.

À la section 5, nous discutons de l'effet résultant des effets combinés de la pondération, des grappes spatiales et de l'intervieweur. La formule de l'effet global donne une détermination exacte de la taille d'échantillon à l'étape de la planification. Nous présentons une belle ventilation de cet effet global en effets dus à la pondération, aux grappes et à l'intervieweur. Cette décomposition peut faciliter la découverte de divers moyens de réduire la variance totale. Dans sa discussion de l'article de Verma, Scott et O'Muircheartaigh (1980), Hedges a mentionné qu'une telle formule de l'effet global était nécessaire. Nous généralisons une formule proposée antérieurement par Davis et Scott (1995) à un plan non EPE et à un modèle général de corrélation valide pour les données discrètes ainsi que continues. Les preuves de tous les résultats techniques sont présentées à l'annexe.

2. Plan EPE sans grappes spatiales

Soit y_k l'observation obtenue auprès de la $k^{\text{ième}}$ personne interrogée par le $i^{\text{ième}}$ intervieweur ($i = 1, \dots, n$). Définissons $n = \sum_{i=1}^n n_i$, la taille totale d'échantillon, $\bar{y} = 1/n \sum_{i=1}^n \sum_{k=1}^{n_i} y_k$, la moyenne d'échantillon non pondérée et $\bar{n}_{int}(\mathbf{a}) = \sum_{i=1}^n a_i n_i$, une moyenne pondérée des charges de travail d'intervieweur, où a_i est un poids arbitraire appliqué à la charge de travail du $i^{\text{ième}}$ intervieweur et $\mathbf{a} = (a_1, \dots, a_i)$.

Nous commençons par donner une justification assistée par modèle de la formule classique d'effet d'intervieweur, c'est-à-dire $ieff = 1 + (\bar{n}_{int} - 1) p_{int}$, où \bar{n}_{int} est la moyenne non pondérée des charges de travail de l'intervieweur. Notons que $\bar{n}_{int} = \bar{n}_{int}(\mathbf{a}_0)$, avec $\mathbf{a}_0 = (a_0, \dots, a_0)$, $a_0 = 1/I$ et $ieff = ieff(\mathbf{a}_0)$. En utilisant le résultat I donné à l'annexe, nous obtenons

$$ieff(\mathbf{a}_1) = \frac{Var_{M_2}(\bar{y})}{Var_{M_1}(\bar{y})} = 1 + [\bar{n}_{int}(\mathbf{a}_1) - 1] p_{int}$$

où $\mathbf{a}_1 = (a_1, \dots, a_i)$, avec $a_i = n_i/n$. Dans la formule susmentionnée, $Var_{M_1}(\bar{y})$ et $Var_{M_2}(\bar{y})$ sont les variances de \bar{y} sous les deux modèles suivants, respectivement,

$$M_1 : Cov(y_k, y_{k'}) = \begin{cases} \sigma^2 & \text{si } i = i', k = k', \\ 0 & \text{autrement,} \end{cases} \quad M_2 : Cov(y_k, y_{k'}) = \begin{cases} \sigma^2 & \text{si } i = i', k = k', \\ p_{int} \sigma^2 & \text{si } i = i', k \neq k', \\ 0 & \text{autrement.} \end{cases}$$

Notons que, contrairement au modèle M_1 , le modèle M_2 introduit l'homogénéité des observations recueillies par un même intervieweur.

Remarque 2.1 : Il découle du corollaire du résultat I, donné à l'annexe, que pour $p_{int} > 0$, $ieff(\mathbf{a}_1) = ieff$ si et uniquement si $n_i = n/I$ pour tout i , c'est-à-dire si et uniquement si la charge de travail est la même pour chaque intervieweur. Pour le cas équilibré, Kish (1962) a donné une justification assistée par modèle de $ieff$ en utilisant un modèle linéaire mixte qui est un cas particulier de M_2 . Pour le cas non équilibré, il est intéressant de souligner la similitude entre la formule de la variabilité d'intervieweur $ieff(\mathbf{a}_1)$ et la formule des effets de plan (A3) donnée dans l'article de Holt discutant de Verma et coll. (1980).

Remarque 2.2 : Il découle du corollaire du résultat I que, si $p_{int} > 0$ et que les n_i ne sont pas égaux, alors $ieff(\mathbf{a}_1) > ieff$.

Dans l'exemple qui suit, nous démontrons la mesure dans laquelle $ieff(\mathbf{a}_1)$ et $ieff$ pourraient différer pour divers profils de charges de travail de différents intervieweurs.

Exemple 1 : Dans le tableau 1, nous considérons trois affectations de charges de travail différentes pour dix intervieweurs, chacune avec $n = 790$. Le cas (A) représente l'affectation des charges de travail la plus variable avec un écart-type = 68,3, le cas (B) est presque équilibré avec un écart-type = 9,5 et le cas (C) correspond à l'affectation de charges de travail égales aux intervieweurs.

Tableau 1

Trois affectations différentes des charges de travail aux intervieweurs (exemple 1)

Profil des charges de travail des intervieweurs		
Intervieweurs		
A)	B)	C)
1	4	79
2	10	79
3	20	79
4	34	79
5	52	79
6	74	79
7	100	79
8	130	79
9	164	79
10	202	79
$\bar{n}_{int}(\mathbf{a}_1)$	790	790

Soit $ieff(\mathbf{a}_{1:A}), ieff(\mathbf{a}_{1:B})$ et $ieff(\mathbf{a}_{1:C}) = ieff$ qui désignent par modèle correspondant aux cas A, B et C, respectivement. Pour $p_{int} > 0$, la fonction $ieff(\mathbf{a}_1)$ est Schur-convexe, ce qui explique que $ieff(\mathbf{a}_{1:A}) \geq ieff(\mathbf{a}_{1:B}) \geq ieff(\mathbf{a}_{1:C}) = ieff$. La figure 1 donne les

intervieweurs est remplacée par la charge de travail moyenne des intervieweurs, c'est-à-dire la formule $1 + (\bar{\pi}_{im} - 1) p_{im}$. À la section 2, nous soutenons que cette formule standard $1 + (\bar{\pi}_{im} - 1) p_{im}$ ne peut pas être interprétée comme un accroissement de la variance totale causé par les intervieweurs, même dans le cas d'un plan EPF avec charges de travail inégales des intervieweurs. Aux sections 2 à 4, nous observons que la définition de la variance d'intervieweur dépend de la nature du plan d'échantillonnage complexe, ainsi que de la charge de travail affectée à l'intervieweur. Dans le présent article, nous donnons des définitions appropriées de la variance d'intervieweur sous divers scénarios de sondage. Une définition fiable de la variance d'intervieweur aide à déterminer les mesures qui doivent être prises afin de réduire la variabilité d'intervieweur. Les résultats présentés dans l'article sont avant tout applicables à la planification des enquêtes plutôt qu'à l'analyse des données d'enquête. Autrement dit, nous nous concentrons ici sur les définitions et sur l'interprétation de la variabilité d'intervieweur et non sur son estimation pour une enquête particulière.

À la section 2, nous considérons un plan EPF sans grappes spatiales et nous fournissons une interprétation assistée par modèle de l'effet d'intervieweur *ieff*. Nous montrons que, pour des charges de travail d'intervieweur égales, *ieff* est simplement égal au ratio de la variance de la moyenne d'échantillon sous un modèle corréle tenant compte de l'homogénéité des observations recueillies par un même intervieweur à celle sous un modèle non corréle qui ne tient pas compte de cette homogénéité. Donc, en multipliant la variance de la moyenne d'échantillon pour l'échantillonnage aléatoire simple par *ieff*, nous pouvons obtenir la variance totale de la moyenne d'échantillon qui tient compte à la fois de la variabilité d'échantillonnage et de la variabilité d'intervieweur. Il s'agit d'une interprétation très intuitive de *ieff* qui complète la justification assistée par modèle donnée antérieurement par Kish (1962). Dans cette section, nous montrons aussi que, pour un plan EPF, *ieff* est plus faible que la valeur donnée par la formule de l'effet d'intervieweur assistée par modèle si la charge de travail de l'intervieweur varie et que la corrélation intra-intervieweur est positive. Donc, le concepteur d'enquêtes qui utilise *ieff* fera moins d'effort qu'il n'est réellement nécessaire pour contrôler la variabilité d'intervieweur. Dans cette situation, une formule appropriée de l'effet d'intervieweur peut être obtenue à partir de *ieff* en substituant la charge de travail moyenne pondérée d'intervieweur à la moyenne simple habituelle.

À la section 3, nous examinons la possibilité d'une pondération inégale, mais sans grappes spatiales. Nous obtenons une interprétation assistée par modèle de *ieff* si et uniquement si les personnes interrogées par un même

intervieweur ont le même poids de sondage et que la charge de travail de l'intervieweur est inversement proportionnelle au carré du poids commun pour l'intervieweur. Il est intéressant de souligner que, contrairement au plan EPF, une répartition égale des charges de travail d'intervieweur ne garantit pas nécessairement une interprétation assistée par modèle de *ieff*. En cas de charges de travail égales des intervieweurs, s'il existe au moins un intervieweur pour lequel les enquêtes n'ont pas tous le même poids de sondage, nous montrons que *ieff* est toujours plus élevé que la valeur donnée par la formule assistée par modèle. Nous dégageons aussi les facteurs à l'origine de la différence entre ces deux formules. Ces résultats ont un intérêt pratique en ce qui concerne la réduction des coûts d'enquête. Plus précisément, le concepteur d'enquêtes qui utilise *ieff* affectera vraisemblablement un budget plus important qu'il n'est vraiment nécessaire au contrôle de la variabilité d'intervieweur. Nous avons également mentionné certaines situations où *ieff* pourrait poser un problème de sous-estimation et, par conséquent, les concepteurs d'enquêtes qui utilisent cette formule pourraient accorder trop peu d'importance au contrôle des effets d'intervieweur. Notre formule fournit une évaluation plus exacte de la variabilité d'intervieweur et permet donc d'affecter un budget plus raisonnable au contrôle de la variabilité d'intervieweur. De surcroît, la modification des formules de planification aura une incidence sur la taille de l'échantillon. Dans de nombreuses enquêtes par sondage à grande échelle, pour diverses raisons de nature organisationnelle et financière, telles que l'absence d'un registre de population général ou la réduction du coût global d'enquête, un plan d'échantillonnage en grappes à plusieurs degrés est considéré comme une alternative moins coûteuse que l'échantillonnage aléatoire simple. Sous un plan d'échantillonnage en grappes à plusieurs degrés, des personnes vivant dans des lieux spatialement rapprochés sont sélectionnées. Les personnes qui vivent dans une même grappe spatiale ont tendance à avoir les mêmes attitudes, à cause de leur contexte socioéconomique semblable, ce qui accroît l'homogénéité interne des données d'enquête. Cette homogénéité spatiale viole l'hypothèse iid (unités indépendantes et identiquement distribuées) fréquemment émise dans les méthodes d'inférence statistique standard, au même titre que la mise en grappes des intervieweurs. Ce fait a été reconnu par de nombreux spécialistes de la recherche sur les enquêtes et l'apport de corrections à diverses procédures statistiques, ainsi que les problèmes logistiques connexes ont été décrits dans la littérature (voir, entre autres, Rao et Scott 1984; Skinner, Holt et Smith 1989; Biemer et Trewin 1997; Chambers et Skinner 2003). À la section 4, nous présentons une nouvelle définition de la variabilité d'intervieweur en présence de pondération inégale et de grappes

De la définition et de l'interprétation de la variabilité d'intervieweur pour un plan d'échantillonnage complexe

Siegfried Gabler et Partha Lahiri¹

Résumé

La variabilité d'intervieweur est une composante importante de la variabilité des statistiques produites par sondage. Diverses stratégies liées au format et à la formulation des questions, ainsi qu'à la formation, à la charge de travail, à l'expérience et à l'affectation des intervieweurs sont employées pour essayer de réduire la variabilité d'intervieweur. La formule classique de mesure de la variabilité d'intervieweur, souvent appelée effet d'intervieweur, est donnée par $teff_{int} = def_{int} - 1 + (\overline{p}_{int} - 1)p_{int}$, où p_{int} et \overline{p}_{int} sont, respectivement, la corrélation intra-intervieweur et la corrélation simple des charges d'échantillon et que les charges de travail des intervieweurs sont égales. Toutefois, les grappes spatiales ainsi que la pondération inégale sont très fréquentes dans les enquêtes à grande échelle. Dans le contexte d'un plan d'échantillonnage complexe, nous obtenons une formule appropriée de la variabilité d'intervieweur qui tient compte des probabilités inégales de sélection et des grappes spatiales. Notre formule fournit une évaluation plus exacte des effets d'intervieweur et permet donc d'affecter un budget plus raisonnable au contrôle de la variabilité d'intervieweur. Nous proposons aussi une décomposition de l'effet global en effets dus à la pondération, aux grappes spatiales et aux intervieweurs. Cette décomposition aide à comprendre différents moyens de réduire la variance totale.

Mots clés : Effet d'intervieweur ; charges de travail des intervieweurs ; corrélation intra-intervieweur ; grappe spatiale ; pondération inégale.

1. Introduction

L'intervieweur est une source importante d'erreurs de mesure dans les enquêtes. Ce fait avait déjà été reconnu par Rice en 1929 et l'a été plus tard par de nombreux spécialistes de la recherche sur les enquêtes. Des facteurs tels que la qualité de la conception du questionnaire et les caractéristiques de l'intervieweur peuvent influencer sur les effets d'intervieweur dans les statistiques produites par sondage.

L'intervieweur peut introduire dans les données d'enquête une homogénéité qui réduit généralement la taille effective de l'échantillon et, par conséquent, augmente la variance totale d'un estimateur. Traditionnellement, l'homogénéité intra-intervieweur a été mesurée au moyen du coefficient de corrélation intra-intervieweur p_{int} . L'importance de ce type de corrélation a été étudiée par de nombreux chercheurs, principalement dans le contexte des enquêtes téléphoniques sans grappes spatiales (Kish 1962; Gray 1956; Hanson et Marks 1958; Tucker 1983; Groves et Magliav 1986; Heeb et Gmel 2001, et d'autres). Ces chercheurs ont soutenu que la nature des questions d'enquête peut avoir une incidence sur la valeur de p_{int} . Les questions sur les attitudes et les questions factuelles complexes sont considérées comme étant plus sensibles à la corrélation intra-intervieweur que les questions factuelles simples (Collins et Butcher 1982; Feather 1973; Fellegi 1964; Gray 1956; Hansen, Hurwitz et Bershad 1961).

Selon Groves (1989), des valeurs supérieures à 0,1 sont

rarement observées. Pour une discussion plus approfondie de la question, voir Schnell et Kreuter (2005). Comme l'ont souligné plusieurs chercheurs, la formule standard de l'effet d'intervieweur $1 + (\overline{p}_{int} - 1)p_{int}$ donne à penser que, même si la corrélation intra-intervieweur est faible, l'effet d'intervieweur pourrait être important, simplement à cause d'une charge moyenne de travail de l'intervieweur élevée. Par exemple, si $p_{int} = 0,01$ et $\overline{p}_{int} = 70$, nous avons $teff = 1,69$ (Schnell et Kreuter 2005). Il convient de souligner qu'une charge de travail moyenne d'intervieweur élevée (par exemple de 60 à 70 cas) est très fréquente dans les enquêtes téléphoniques (Tucker 1983; Groves et Magliav 1986). Dans le cas de l'Enquête sociale européenne, Philipps et Loosveldt (2004) ont produit des boîtes à moustaches des corrélations intra-intervieweur et des charges de travail d'intervieweur pour 18 pays participants.

L'effet, ou variance, d'intervieweur est généralement défini comme étant l'accroissement de la variance totale causée uniquement par les intervieweurs. Pour un plan EPF avec charges de travail égales des intervieweurs, la variance d'intervieweur pour la moyenne d'échantillon est donnée simplement par $1 + (\overline{p}_{int} - 1)p_{int}$, où p_{int} est la charge de travail commune des intervieweurs. Dans le cas des enquêtes complexes avec charges de travail inégales des intervieweurs, les spécialistes de la recherche sur les enquêtes utilisent fréquemment une modification simple de cette formule où la charge de travail commune des intervieweurs est donnée

1. Siegfried Gabler, GESIS, C.P. 12 21 55, 68072 Mannheim, Allemagne. Courriel : siegfried.gabler@gesis.org; Partha Lahiri, Université du Maryland, College Park, États-Unis. Courriel : plahiri@survey.umd.edu.

- Maravall, A., et Caporello, G. (2004). Program TSW: Revised Reference Manual. Document de travail 2004, Research Department, Bank of Spain. <http://www.bde.es>.
- Nerlove, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica*, 32, 241-286.
- Newton, H., et Pagano, M. (1983). A method for determining periods in time series. *Journal of the American Statistical Association*, 78, 152-157.
- Parzen, E. (1983). Autoregressive spectral estimation. *Handbook of Statistics III*, (Éds., D. Brillinger et P. Krishnaiah), Amsterdam : North Holland, 221-247.
- Priestley, M. (1981). *Spectral Analysis and Time Series*. London : Academic Press.
- Soukup, R.J., et Findley, D.F. (1999). On the spectrum diagnostics used by X-12-ARIMA to indicate the presence of trading day effects after modeling or adjustment. Aussi au www.census.gov/pub/ts/papers/t9903s.pdf. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 144-149.
- Taniguchi, M., et Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. New York City, New York : Springer-Verlag.
- U.S. Census Bureau (2002). X-12 ARIMA Reference Manual (Version 0.2.10), Washington, DC.

Théorème 1 Supposons que les cumulants de quatrième ordre de $\{X_t\}$ disparaissent ; que la condition (B) ou (HT) est vérifiée ; et que l'hypothèse 1 (8) de Chiu (1988) est vérifiée. Soit les noyaux A et B satisfaisant les conditions (i) à (iv) de la section 2.1. Alors

$$\left\{ \sqrt{n} \frac{(\theta^A(I) - \theta^A(J))}{\sqrt{\theta^A(I^2) - \theta^A(J^2)}} \right\} \xrightarrow{\mathcal{D}} N(0, V)$$

quand $n \rightarrow \infty$. Ici, 0 dénote le vecteur nul $(0, 0)$, et V est une matrice de dimension 2×2 dont les entrées sont

$$V_{11} = V_{22} = 1 \quad V_{12} = V_{21} = \frac{\theta^A(J^2)(f_2^2)\theta^B(J^2)(f_2^2)}{\theta^B(J^2)(f_2^2)}.$$

Preuve. Pour commencer, établissons que $\theta^A(I^2) \xrightarrow{P} \theta^A(J^2)$

$2\theta^A(J^2)$. Puisque le noyau A est continu dans un intervalle (tel que $[\mu - \beta/2, \mu + \beta/2]$, ce résultat découle directement du corollaire 1 de Chiu (1988), en notant qu'il traite l'approximation de la fonction intégrale par les sommes de Riemann (Chiu (1988) définit aussi le pétiodogramme avec un facteur 2π). Évidemment, les mêmes résultats tiennent avec B à la place de A . Deuxièmement, considérons la convergence conjointe de $\theta^A(I)$ et $\theta^B(I)$. Nous utilisons la technique de Cramér-Wold et appliquons le lemme 3.1.1 de Taniguchi et Kakizawa (2000), généralisé comme il convient pour inclure les fonctions impaires (voir le théorème 3 de Chiu (1988)). Donc, pour tout x, y réel,

$$\sqrt{n} \left[x \frac{(\theta^A(I) - \theta^A(J))}{(\theta^B(I) - \theta^B(J))} + y \frac{\sqrt{\theta^A(J^2)(f_2^2)}}{\sqrt{\theta^B(J^2)(f_2^2)}} \right] \xrightarrow{\mathcal{D}} N \left(0, \frac{1}{2\pi} \int_{-\pi}^{\pi} (C(\lambda)C(-\lambda) + C^2(\lambda))(f_2^2(\lambda))d\lambda \right)$$

en utilisant le théorème de Slutsky (Bickel et Doksum 1977), où le noyau C est défini par

$$C(\lambda) = \frac{A(\lambda)}{x} \frac{\sqrt{\theta^A(J^2)(f_2^2)}}{\sqrt{\theta^B(J^2)(f_2^2)}} B(\lambda).$$

De toute évidence, $C(-\lambda) = 0$ et

$$C^2(\lambda) = \frac{\theta^A(J^2)(f_2^2)}{x^2} A^2(\lambda)$$

$$+ 2 \frac{\sqrt{\theta^A(J^2)(f_2^2)} \sqrt{\theta^B(J^2)(f_2^2)}}{xy} A(\lambda) B(\lambda) + \frac{\theta^B(J^2)(f_2^2)}{y^2} B^2(\lambda).$$

Statistique Canada.

Lothian, J. et Morry, M. (1978). A test for identifiable seasonality when using the X-11-ARIMA program. Document de travail, Division de la recherche et analyse en séries chronologiques, Statistique Canada.

Hosoya, Y., et Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics*, 10, 132-153.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. et Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177 (avec discussion).

Findley, D. (2006). Communication personnelle. Association, [CD-ROM] : Alexandria, VA).

Evans, T., Holan, S. et McElroy, T. (2006). Evaluating measures for assessing spectral peaks. *2006 Proceedings American Statistical Association*, 16, 1315-1326.

Chiu, S. (1988). Weighted least squares estimators on the frequency domain for the parameters of a time series. *The Annals of Statistics*, 16, 1315-1326.

Francisco : Holden-Day. Brillinger, D. (1981). *Time Series Data Analysis and Theory*. San Francisco.

Benjamin, Y., et Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Bickel, P., et Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Englewood Cliffs, New Jersey : Prentice Hall.

Bell, W., et Hillmer, S. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-320.

Aston, J., Findley, D., McElroy, T., Willis, K. et Martin, D. (2007). New ARIMA Models for Seasonal Time Series and Their Application to Seasonal Adjustment and Forecasting. *SRD Research Report No. RRS 2007-14*, U.S. Census Bureau.

Bibliographie

multiples H-FWER.

Dans ce cas, si le support pour n importe quelle paire de noyaux est disjoint, nous obtenons l'indépendance asymptotique et nous pouvons par conséquent appeler la procédure de tests

$$\frac{\sqrt{\theta^{A_{i'}}(f_2^2)\theta^{A_{j'}}(f_2^2)}}{\theta^{A_{i'}}(f_2^2)}.$$

asymptotique V aura la i^{e} entrée

d noyaux comme il suit. La matrice de covariance d noyaux généraliser le théorème 1 pour passer de deux à chacun les hypothèses de la section 2. Alors, nous pouvons finir de noyaux A_i pour $i = 1, 2, \dots, d$, qui satisfont une de pics multiples. Supposons donc que nous avons une série

Ensuite, nous discutons du scénario de test de détection combinés. En prenant x et y égaux à zéro et un dans diverses combinaisons, nous déduisons la matrice de variance

Tableau 10

Analyses des données pour 17 séries britanniques de l'OCDE comparant notre diagnostic de pics multiples s'appuyé sur la méthode H-F-WER pour contrôler le taux de erreur de type I global à $\alpha = 0,05$ et $\alpha = 0,10$ et examine la pente à $\delta = 0,10$ (voir la section 4.2)

M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-F-WER pour contrôler le taux de erreur de type I global à $\alpha = 0,05$ et $\alpha = 0,10$ et examine la pente à $\delta = 0,10$ (voir la section 4.2)

Analyses des données – OCDE, Grande-Bretagne											
Série	Données originales						Données désaisonnalisées				
	H-FWER 0,10/0,05			H-FWER 0,10/0,10			H-FWER 0,10/0,05			H-FWER 0,10/0,10	
	TH	Quartique	TH	TH	Quartique	TH	TH	Quartique	TH	TH	SV
BRETAGNE											
PPIAMP01	1234	1234	1234	1234	1234	24	0,56	1,58	0,5	0,5	0,5
PPIAMP02	1	1	123	123	2	0,53	0,92	0,5	0,5	0,5	0,5
PPPFU01	1345	12345	1345	12345	12	0,64	0,59	0,5	0,5	0,5	0,5
PRINT001	1345	1345	1345	1345	23	0,16	0,40	0,5	0,5	0,5	0,5
PRMNCG02	2345	2345	2345	2345	123	0,23	0,56	0,5	0,5	0,5	0,5
PRMNCG03	12345	12345	12345	12345	123	0,20	0,49	0,5	0,5	0,5	0,5
PRMNC501	123	12	123	1234	12	0,68	1,31	0,5	0,5	0,5	0,5
PRMNIG01	2345	2345	2345	2345	123	0,15	0,47	1	0,5	0,5	0,5
PRMNT001	1345	1345	1345	1345	23	0,17	0,45	0,5	0,5	0,5	0,5
PRMNV002	12345	12345	12345	12345	12345	0,25	0,76	0,5	0,5	0,5	0,5
PRMNV003	1234	1234	1234	1234	234	0,29	0,91	0,5	0,5	0,5	0,5
PRMNV001	124	134	124	134	234	0,18	0,58	0,5	0,5	0,5	0,5
SLRTRC03	1345	1345	1345	1345	124	0,42	0,74	124	123	124	123
SLRTRT02	12345	12345	12345	12345	12345	0,05	0,15	1234	0,5	1234	0,5
UNLVRG01	12345	12345	12345	12345	1245	0,63	0,61	0,5	0,5	0,5	0,5
XTEXVA01	134	134	1345	1345	23	0,34	1,02	0,5	0,5	0,5	0,5
XTMVA01	23	23	23	23	23	0,31	0,90	0,5	0,5	0,5	0,5
Total de résultats	17/17	17/17	17/17	17/17	17/17	17/17	14/17	14/17	16/17	13/17	15/17
« corrects »											
Nombre moyen	3,76	3,76	3,94	4,06	2,76	0,47	0,18	0,53	0,47	0	0

Remerciements

Le présent article est diffusé en vue de tenir les parties intéressées au courant de la recherche courante et de favoriser la discussion des travaux en cours. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau. Les travaux de Holian ont été financés par une bourse de recherche de l'ASA/NSF/BLS.

Annexe

Le présent article est diffusé en vue de tenir les parties intéressées au courant de la recherche courante et de favoriser la discussion des travaux en cours. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau. Les travaux de Holan ont été financés par une bourse de recherche de l'ASA/NSF/BLS.

Remerciements

(B), due à Brillinger (1981), énonce que le processus est strictement stationnaire et la condition (B1) de Taniguchi et Kakizawa (2000, page 55) est vérifiée. La condition (HT), due à Hosoya et Taniguchi (1982), énonce que le processus a une représentation linéaire et les conditions (H1) à (H6) de Taniguchi et Kakizawa (2000, pages 55-56) sont vérifiées. L'hypothèse 1 (8) de Chiu (1988) est une condition de sommabilité sur divers cumulants d'ordre élevé, qui est satisfait, par exemple, par un processus gaussien dont la densité spectrale est dans C^2 . Aucune de ces conditions n'est stricte ; par exemple, un processus causal linéaire avec quatre moments satisfait (HT). Le résultat principal est une convergence conjointe de toute paire de mesures $\psi_A(I)$; par exemple, il peut s'agir d'une mesure de pente et de convexité avec le même noyau A . Nous présentons le théorème général qui couvre ces deux cas.

Tableau 8 (suite)
Analyses des données pour 15 séries sur la zone euro de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0,05$ et $\alpha = 0,10$ et examine la pente à $\delta = 0,10$ (voir la section 4.2)

Analyses des données – OCDE, Euro														
Série	Données originales					Données désaisonnalisées					SV			
	TH	Quartique	TH	Quartique	TH	TH	Quartique	TH	Quartique	TH		TH	Quartique	TH
SLMNIG02	12345	12345	12345	12345	12345	23	0,21	0,45	0,21	0,45	0,45	0,45	0,45	0,45
SLMNT002	1345	2345	1345	12345	12345	23	0,20	0,41	24	0,41	24	24	24	34
SLMNVG02	12345	12345	12345	12345	12345	2345	0,17	0,30	1	0,30	1	1	1245	1345
SLRRT001	12345	12345	12345	12345	12345	0,05	0,12	0,05	0,12	0,05	0,12	0,05	0,12	0,05
SLRRT002	12345	12345	12345	12345	12345	12345	0,05	0,11	0,05	0,11	0,05	0,11	0,05	0,11
XTEXVA01	1345	2345	1345	2345	1345	23	0,31	0,57	0,31	0,57	0,31	0,57	0,31	12
XTMVA01	2345	2345	2345	2345	2345	23	0,40	0,72	0,40	0,72	0,40	0,72	0,40	0,72
Total de résultats	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15	15/15
« corrects »	4,40	4,40	4,40	4,40	4,40	3,73	3,73	3,73	3,73	3,73	3,73	3,73	3,73	3,73
Nombre moyen	4,40	4,40	4,40	4,40	4,40	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20

Tableau 9
Analyses des données pour 11 séries françaises de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0,05$ et $\alpha = 0,10$ et examine la pente à $\delta = 0,10$ (voir la section 4.2)

Analyses des données – OCDE, France														
Série	Données originales					Données désaisonnalisées					SV			
	TH	Quartique	TH	Quartique	TH	TH	Quartique	TH	Quartique	TH		TH	Quartique	TH
PRAFA001	12345	12345	12345	12345	12345	123	0,13	0,29	0,13	0,29	0,29	0,29	0,29	0,29
PRNTO001	2345	2345	2345	2345	2345	235	0,14	0,44	0,14	0,44	0,44	0,44	0,44	0,44
PRMNCG01	12345	12345	12345	12345	12345	234	0,15	0,38	1	0,38	1	1	1	1
PRMNCG01	12345	12345	12345	12345	12345	234	0,25	0,58	0,25	0,58	0,58	0,58	0,58	0,58
PRMNG01	12345	12345	12345	12345	12345	12345	0,11	0,26	0,11	0,26	0,26	0,26	0,26	0,26
PRMNT001	12345	12345	12345	12345	12345	12345	0,16	0,29	123	0,29	123	123	123	123
PRMNV001	12345	12345	12345	12345	12345	12345	0,24	0,34	0,24	0,34	0,34	0,34	0,34	0,34
SLRTR001	1345	2345	1345	2345	1345	123	0,27	0,71	0,27	0,71	0,71	0,71	0,71	0,71
SLRRT002	12345	12345	12345	12345	12345	12345	0,16	0,36	0,16	0,36	0,36	0,36	0,36	0,36
XTEXVA01	1345	1345	1345	1345	1345	23	0,14	0,44	0,14	0,44	0,44	0,44	0,44	0,44
XTMVA01	1245	1245	1245	1245	1245	23	0,18	0,54	0,18	0,54	0,54	0,54	0,54	0,54
Total de résultats	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11
« corrects »	4,64	4,64	4,64	4,64	4,64	3,55	3,55	3,55	3,55	3,55	3,55	3,55	3,55	3,55
Nombre moyen	4,64	4,64	4,64	4,64	4,64	0,36	0,36	0,36	0,36	0,36	0,36	0,36	0,36	0,36

Tableau 6
Analyses des données pour 30 séries du U.S. Census Bureau (10 sur le logement, 10 sur les importations/exportations et 10 sur les ventes au détail) comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0,05$ et $\alpha = 0,10$ et examine la perte à $8 = 0,10$ (voir la section 4.2).

Analyses des données – Séries sur la fabrication												
Série	Quartique	TH	Données originales			Données désaisonnalisées			Série	Quartique	TH	SV
			H-FWER 0,10/0,05	H-FWER 0,10/0,10	SV	M7	M8	H-FWER 0,10/0,05				
MWIFam	12345	12345	12345	12345	12	0,13	0,25	12345	12345	12345	12345	12345
NWToi	12345	12345	12345	12345	12	0,18	0,31	12345	12345	12345	12345	12345
NEIFam	12345	12345	12345	12345	12	0,16	0,33	12345	12345	12345	12345	12345
NEToi	12345	12345	12345	12345	123	0,25	0,27	12345	12345	12345	12345	12345
SIFam	12345	12345	12345	12345	125	0,22	0,47	12345	12345	12345	12345	12345
SToi	124	124	1245	1245	125	0,29	0,57	12345	12345	12345	12345	12345
USIFam	12345	12345	12345	12345	125	0,17	0,39	12345	12345	12345	12345	12345
USToi	12345	12345	12345	12345	125	0,20	0,42	12345	12345	12345	12345	12345
WIFam	1234	1234	12345	12345	125	0,21	0,44	12345	12345	12345	12345	12345
WToi	1234	1234	12345	12345	12	0,27	0,56	12345	12345	12345	12345	12345
Total des résultats	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Séries sur les importations/exportations												
Série	Quartique	TH	Données originales			Données désaisonnalisées			Série	Quartique	TH	SV
			H-FWER 0,10/0,05	H-FWER 0,10/0,10	SV	M7	M8	H-FWER 0,10/0,05				
M00120	12345	12345	12345	12345	125	0,23	0,48	12345	12345	12345	12345	12345
M00190	12345	12345	12345	12345	1235	0,38	0,59	12345	12345	12345	12345	12345
M3000	12345	12345	12345	12345	234	0,48	0,95	12345	12345	12345	12345	12345
M3010	1234	1234	12345	12345	2345	0,52	0,88	12345	12345	12345	12345	12345
X3	12345	12345	12345	12345	2345	0,57	0,94	12345	12345	12345	12345	12345
X00300	134	134	134	134	2	0,56	0,97	12345	12345	12345	12345	12345
X3020	12345	12345	12345	12345	12345	0,39	0,70	12345	12345	12345	12345	12345
X3022	12345	12345	12345	12345	23	0,69	1,04	12345	12345	12345	12345	12345
X10140	1234	1234	1234	1234	15	0,29	0,47	1234	1234	1234	1234	1234
Total des résultats	10/10	9/10	10/10	10/10	10/10	10/10	9/10	10/10	10/10	10/10	10/10	10/10
Séries sur les ventes de détail												
Série	Quartique	TH	Données originales			Données désaisonnalisées			Série	Quartique	TH	SV
			H-FWER 0,10/0,05	H-FWER 0,10/0,10	SV	M7	M8	H-FWER 0,10/0,05				
s0b441x0	12345	12345	12345	12345	135	0,22	0,41	12345	12345	12345	12345	12345
s0b 44000	12345	12345	12345	12345	2345	0,12	0,26	12345	12345	12345	12345	12345
s0b 44100	12345	12345	12345	12345	135	0,21	0,40	12345	12345	12345	12345	12345
s0b 44130	12345	12345	12345	12345	1235	0,21	0,42	12345	12345	12345	12345	12345
s0b 44200	12345	12345	12345	12345	12345	0,13	0,27	12345	12345	12345	12345	12345
s0b 44300	1234	12345	1234	12345	12345	0,12	0,18	12345	12345	12345	12345	12345
s0b 44312	1234	1234	1234	1234	1234	0,31	0,48	12345	12345	12345	12345	12345
s0b 44400	12345	12345	12345	12345	1235	0,16	0,32	1235	0,16	0,32	1235	0,16
s0b 44410	12345	12345	12345	12345	12345	0,14	0,22	1235	0,14	0,22	1235	0,14
s0b 44500	12345	12345	12345	12345	12345	0,23	0,33	1235	0,23	0,33	1235	0,23
Total des résultats	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
« corrects »												
Données originales												
Série	Quartique	TH	Données originales			Données désaisonnalisées			Série	Quartique	TH	SV
			H-FWER 0,10/0,05	H-FWER 0,10/0,10	SV	M7	M8	H-FWER 0,10/0,05				
s0b441x0	12345	12345	12345	12345	135	0,22	0,41	12345	12345	12345	12345	12345
s0b 44000	12345	12345	12345	12345	2345	0,12	0,26	12345	12345	12345	12345	12345
s0b 44100	12345	12345	12345	12345	135	0,21	0,40	12345	12345	12345	12345	12345
s0b 44130	12345	12345	12345	12345	1235	0,21	0,42	12345	12345	12345	12345	12345
s0b 44200	12345	12345	12345	12345	12345	0,13	0,27	12345	12345	12345	12345	12345
s0b 44300	1234	12345	1234	12345	12345	0,12	0,18	12345	12345	12345	12345	12345
s0b 44312	1234	1234	1234	1234	1234	0,31	0,48	12345	12345	12345	12345	12345
s0b 44400	12345	12345	12345	12345	1235	0,16	0,32	1235	0,16	0,32	1235	0,16
s0b 44410	12345	12345	12345	12345	12345	0,14	0,22	1235	0,14	0,22	1235	0,14
s0b 44500	12345	12345	12345	12345	12345	0,23	0,33	1235	0,23	0,33	1235	0,23
Total des résultats	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
« corrects »												
Données désaisonnalisées												
Série	Quartique	TH	Données désaisonnalisées			Données désaisonnalisées			Série	Quartique	TH	SV
			H-FWER 0,10/0,05	H-FWER 0,10/0,10	SV	M7	M8	H-FWER 0,10/0,05				
s0b441x0	12345	12345	12345	12345	135	0,22	0,41	12345	12345	12345	12345	12345
s0b 44000	12345	12345	12345	12345	2345	0,12	0,26	12345	12345	12345	12345	12345
s0b 44100	12345	12345	12345	12345	135	0,21	0,40	12345	12345	12345	12345	12345
s0b 44130	12345	12345	12345	12345	1235	0,21	0,42	12345	12345	12345	12345	12345
s0b 44200	12345	12345	12345	12345	12345	0,13	0,27	12345	12345	12345	12345	12345
s0b 44300	1234	12345	1234	12345	12345	0,12	0,18	12345	12345	12345	12345	12345
s0b 44312	1234	1234	1234	1234	1234	0,31	0,48	12345	12345	12345	12345	12345
s0b 44400	12345	12345	12345	12345	1235	0,16	0,32	1235	0,16	0,32	1235	0,16
s0b 44410	12345	12345	12345	12345	12345	0,14	0,22	1235	0,14	0,22	1235	0,14
s0b 44500	12345	12345	12345	12345	12345	0,23	0,33	1235	0,23	0,33	1235	0,23
Total des résultats	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
« corrects »												

Tableau 5
Analyses des données pour 35 séries sur la fabrication (U.S. Census Bureau) comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0,05$ et $\alpha = 0,10$ et examine la pente à $\delta = 0,10$ (voir la section 4.2)

Analyses des données – Séries sur la fabrication											
Série	Données originales			Données désaisonnalisées							
	H-FWER 0,10/0,05	TH	Quartique	H-FWER 0,10/0,05	TH	Quartique	M8	M7	SV	TH	SV
M ₁	2345	1235	2345	1234	1234	0,39	0,24	0,24	12	0,24	0,39
M ₂	12345	12345	12345	12345	1235	0,20	0,32	0,32	1235	0,20	0,32
M ₃	12345	12345	12345	12345	12345	0,28	0,46	0,46	1235	0,28	0,46
M ₄	∅	∅	12345	12345	12345	0,28	0,44	0,44	12	0,28	0,44
M ₅	12345	12345	12345	12345	12345	0,27	0,47	0,47	12345	0,27	0,47
M ₆	∅	∅	123	123	123	0,28	0,49	0,49	12	0,28	0,49
M ₇	12345	12345	12345	12345	12345	0,18	0,37	0,37	12345	0,18	0,37
M ₈	12345	12345	12345	12345	12345	0,42	0,73	0,73	124	0,42	0,73
M ₉	∅	∅	12345	12345	12345	0,38	0,72	0,72	1	0,38	0,72
M ₁₀	∅	∅	∅	1234	1234	0,15	0,27	0,27	123	0,15	0,27
M ₁₁	1234	1234	1234	1234	1234	0,30	0,54	0,54	1234	0,30	0,54
M ₁₂	1234	1234	1234	1234	1234	0,24	0,39	0,39	1234	0,24	0,39
M ₁₄	∅	∅	1234	1234	1234	0,24	0,40	0,40	1234	0,24	0,40
M ₁₅	12345	12345	12345	12345	12345	0,23	0,43	0,43	12345	0,23	0,43
M ₁₆	1234	1234	1234	1234	1234	0,23	0,40	0,40	1234	0,23	0,40
M ₁₇	∅	∅	1234	1234	1234	0,64	0,66	0,66	12	0,64	0,66
M ₁₈	12345	12345	12345	12345	12345	0,20	0,37	0,37	245	0,20	0,37
M ₁₉	∅	∅	∅	∅	12345	1,00	∅	∅	4	0,86	1,00
M ₂₀	12345	12345	12345	12345	12345	0,56	0,84	0,84	4	0,56	0,84
M ₂₁	12345	12345	12345	12345	12345	0,37	0,58	0,58	1234	0,37	0,58
M ₂₂	12345	12345	12345	12345	12345	0,26	0,45	0,45	1234	0,26	0,45
M ₂₃	12345	12345	12345	12345	12345	0,20	0,47	0,47	1234	0,20	0,47
M ₂₄	12345	12345	12345	12345	12345	0,26	0,43	0,43	2345	0,26	0,43
M ₂₅	12345	12345	12345	12345	12345	0,27	0,42	0,42	12345	0,27	0,42
M ₂₆	12345	12345	12345	12345	12345	0,37	0,62	0,62	1235	0,37	0,62
M ₂₇	1345	1234	1345	1345	1234	0,25	0,22	0,22	2345	0,25	0,22
M ₂₈	∅	∅	∅	∅	∅	0,57	0,44	0,44	24	0,57	0,44
M ₂₉	12345	12345	12345	12345	12345	0,78	1,13	1,13	24	0,78	1,13
M ₃₀	123	1234	12345	12345	12345	0,45	0,65	0,65	245	0,45	0,65
M ₃₁	∅	∅	123	123	123	0,64	0,46	0,46	4	0,64	0,46
M ₃₂	1235	12345	1235	12345	12345	0,21	0,37	0,37	12345	0,21	0,37
M ₃₃	12345	12345	12345	12345	1234	0,24	0,38	0,38	1234	0,24	0,38
M ₃₄	12345	12345	12345	12345	12345	0,46	0,85	0,85	234	0,46	0,85
M ₃₅	12345	12345	12345	12345	12345	0,25	0,66	0,66	2345	0,25	0,66
M ₃₆	12345	12345	12345	12345	12345	1,32	1,56	1,56	123	1,32	1,56
Total de résultats											
« corrects »											
Nombre moyen	3,09	3,29	4,09	4,09	3,23	0	0	0	0	0,26	0

Pour le scénario de test de détection de pics multiples, nous employons des résultats connus extraits de la littérature sur les tests multiples (c'est-à-dire les applications pour contrôler le taux global d'erreur de type I) pour combiner les valeurs p issues des cinq fréquences saisonnières de manière à accroître considérablement la puissance statistique, comme il est démontré au tableau 4. Malgré un certain écart dans la taille (tableau 3) pour le test de détection de pics multiples, les résultats demeurent relativement utilisables. Sur un lot type de séries saisonnières, le nombre d'erreurs de type I est celui auquel on s'attendrait et la puissance est assez décente (tableaux 5 à 10). Bien que notre méthode se compare assez favorablement aux diagnostics SV, M7 et M8, aucun des diagnostics ne fournit une valeur p et seul le premier permet de discerner quels pics spectraux contribuent au comportement saisonnier. Il s'agit d'un aspect important pour le désaisonnalisateur qui veut savoir non seulement s'il existe une saisonnalité résiduelle, mais aussi à quelles fréquences saisonnières, afin de pouvoir prendre les mesures appropriées pour modifier les filtres de désaisonnalisation (ce qui peut se faire en lissant sur des années supplémentaires, en changeant les filtres saisonniers dans X-11-ARIMA; ou bien, on pourrait envisager de rechercher courants sur une approche fondée sur un modèle racourcir la série. Pour une description des travaux de recherche courants sur une approche fondée sur un modèle pour concevoir des filtres de désaisonnalisation ciblés sur des fréquences saisonnières particulières, voir Aston, Findley, McElroy, Willis et Martin (2007)).

Le choix du noyau à certainement une incidence sur les résultats, quoiqu'en pratique, nous ne constatons que peu de différence entre les noyaux quartique et TH; ce dernier pourrait être marginalement plus puissant. Évidemment, de nombreux autres noyaux d'usage répandu pourraient aussi être utilisés par un praticien et nous en avons seulement choisi deux qui semblaient intuitifs et simples à mettre en œuvre. Le choix de l'emplacement μ est clairement dicté par la caractérisation de la saisonnalité. Puisque la puissance statistique diminue généralement avec β , nous recommandons systématiquement de prendre la valeur maximale de β de sorte que les supports de noyaux soient disjoints, ce qui garantit la propriété d'indépendance asymptotique des divers diagnostics qui est un élément crucial de notre méthode

Le test SV indique que toutes les séries brutes sont saisonnières et que la plupart des séries désaisonnalisées n'ont aucun pic spectral saisonnier; les diagnostics M7 et M8 n'indiquent pas quels pics saisonniers sont présents dans les données brutes. Notre procédure révèle quelques cas (quand $\alpha = 0,10$ pour les tests de convexité) où la correction peut être inadéquate, mais ces cas sont compris dans la fourchette de la proportion attendue d'erreurs de type I. Pour les séries brutes, la puissance empirique (c'est-à-dire la proportion totale de résultats corrects) pour notre méthode varie de 0,66 à 0,89, avec une puissance plus élevée pour le niveau $\alpha = 0,10$, comme il fallait s'y attendre. Dans de nombreux cas, les pics indiqués sont les mêmes que ceux détectés par le test SV, mais parfois ils sont assez différents. Notons que le nombre moyen de pics détectés pour les séries brutes est habituellement beaucoup plus élevé dans le cas de notre procédure que dans celui de la méthode SV, qui produit souvent un nombre moyen de l'ordre de 3,2. Lorsque le seuil α passe de 0,05 à 0,10, notre méthode produit naturellement un nombre moyen plus élevé de pics détectés; le test SV ne peut pas être ajusté de cette façon. Inversement, pour les données désaisonnalisées, le nombre moyen de pics détectés a tendance à être inférieur à 1 dans le cas de notre méthode (sauf pour les séries allemandes).

Les résultats sont assez semblables pour les noyaux quartique et TH. Bien que les diagnostics M7, M8 et SV donnent des résultats un peu meilleurs que notre procédure de détection des pics spectraux pour $\alpha = 0,10$, il est important de souligner que notre méthode fournit un niveau de détail que les tests M7 et M8 ne peuvent pas reproduire, tandis que le diagnostic SV ne fournit de valeur p pour aucun des pics (M7 ou M8 ne le font pas non plus). Dans l'ensemble, nous pensons que les résultats sont très encourageants et informatifs.

5. Conclusion

Le présent article décrit une approche novatrice de détection statistique des pics spectraux. Le diagnostic de convexité correspond au calcul d'une moyenne du périodogramme pondérée par la dérivée seconde d'un noyau typique, tel que la fenêtre des décalages de Tukey-Hanning. Implicitement, ce type de statistique comprend une comparaison d'une moyenne du périodogramme près d'une fréquence donnée à sa moyenne un peu plus en dehors; cela découle de la forme générale de $f_{\beta, \mu}$. Le diagnostic de pente aide à décrire les cas où la convexité est négative, mais où il existe aussi une augmentation/diminution importante dans le spectre. Le fait que la méthode fonctionne effectivement comme prévu est prouvé par les simulations et les résultats d'analyse présentés dans les tableaux 1 à 10.

Le présent article décrit une approche novatrice de détection statistique des pics spectraux. Le diagnostic de convexité correspond au calcul d'une moyenne du périodogramme pondérée par la dérivée seconde d'un noyau typique, tel que la fenêtre des décalages de Tukey-Hanning. Implicitement, ce type de statistique comprend une comparaison d'une moyenne du périodogramme près d'une fréquence donnée à sa moyenne un peu plus en dehors; cela découle de la forme générale de $f_{\beta, \mu}$. Le diagnostic de pente aide à décrire les cas où la convexité est négative, mais où il existe aussi une augmentation/diminution importante dans le spectre. Le fait que la méthode fonctionne effectivement comme prévu est prouvé par les simulations et les résultats d'analyse présentés dans les tableaux 1 à 10.

«recolorer» les données comme il est décrit à la section 3.4.

Enfin, les résultats asymptotiques requièrent que les données soient différenciées jusqu'à la stationnarité. Comme les séries chronologiques économiques sont habituellement non stationnaires, il est souhaitable de différencier la tendance des données désaisonnalisées avant d'appliquer notre diagnostic. Cette différenciation pouvant atténuer la détection du premier pic saisonnier, les praticiens peuvent

Pour évaluer la taille, nous avons considéré un test basé sur la convexité seulement ainsi qu'un test basé simultanément sur la pente et la convexité. Les tests portant sur la convexité seulement (C) sont exécutés aux seuils de signification nominaux α de 0,05 et 0,10, en utilisant la méthode H-FWER pour contrôler le taux d'erreur de type I global. Les tests fondés simultanément sur la pente et la convexité (S, C) ont été exécutés de la façon suivante :

1. Exécutez les tests multiples de convexité, $H_0^{(2)}$, en utilisant la méthode H-FWER pour contrôler le taux d'erreur global au niveau α (qui est égal à 0,05 ou 0,10).

2. Pour tout pic significatif découvert à l'étape 1, exécutez le test de pente individuel, $H_0^{(1)}$, au niveau δ (qui est égal à 0,10 ou 0,25). Notons qu'ici, nous souhaitons ne pas rejeter $H_0^{(1)}$ afin de déclarer qu'il n'importe quel « pic » comme étant statistiquement significatif.

3. Déclarer qu'il existe un pic statistiquement significatif si l'étape 1 permet de découvrir une fréquence saisonnière dont la convexité agrégée dans le spectre est significative et que l'étape 2 n'aboutit pas simultanément à la découverte d'une pente agrégée significative pour la fréquence saisonnière correspondante.

Les résultats de cette simulation sont résumés au tableau 3. Un aspect de cette procédure qui nécessite une explication plus approfondie est l'étape 2, où δ (le niveau pour le test de pente) est pris égal à 0,10 et à 0,25. Bien que la partie test de pente de la procédure soit exécutée pour des pics individuels, il semble raisonnable de vouloir être prudent. La question est que, même si certains tests de pente individuels sont rejetés, nous pouvons encore poursuivre dans d'autres cas. Donc la situation observée ici diffère de l'hypothèse classique d'« absence de pics » qui peut être

rejetée si un seul pic est découvert. Évidemment, puisque nous effectuons chaque test d'hypothèse de pente sur la base de pics individuels, tout niveau δ supérieur à 0,05 sera considéré comme plus modéré.

Si, dans le cas de la procédure combinée (S, C), nous ne pouvons pas nous attendre à ce que la taille s'approche de la valeur nominale (parce que l'utilisation du test de pente déséquilibre le taux d'erreur de type I), cette taille n'est pas extrêmement exacte non plus dans le cas de l'utilisation du test de convexité seulement (C), comme le montre l'examen du cas $\alpha = 0,10$ avec $n = 288$, 360. Ici, pour le noyau quantique, la convexité est de trop grande taille, tandis que dans le cas d'un pic unique, le test de convexité possède une taille exacte (tableau 1) pour ces tailles d'échantillon. Notons que la méthode H-FWER produit uniquement une procédure approximativement correctement dimensionnée ; un autre facteur est que les tests pour les cinq pics ne sont qu'asymptotiquement indépendants. Toutes ces raisons font que la taille empirique observée au tableau 3 diffère quelque peu des niveaux nominaux.

Pour étudier la puissance, nous avons considéré la même procédure en trois étapes décrite plus haut. Cependant, pour cette simulation, nous n'avons envisagé que le test conjoint pente-convexité et examiné quatre paires de niveaux (0,10) et (0,25, 0,10). Les résultats de cette simulation (tableau 4) révèlent une puissance énorme, même pour des tailles d'échantillon aussi petites que $n = 120$. Il s'agit d'une propriété extrêmement importante, car $n = 120$ est représentatif des tailles d'échantillon observées en pratique lorsque l'on procède à la désaisonnalisation (par exemple données mensuelles couvrant 10 années). Pour des échantillons de taille $n = 144$, une puissance supérieure à 90 % est réalisée.

Tableau 3 Résultats de la simulation de la taille pour le diagnostic de pics multiples. Ici, 10 000 répétitions ont été utilisées. Le test de convexité a été étudié séparément avec le taux d'erreur de type I global contrôlé à $\alpha = 0,05$ et à $\alpha = 0,10$ en utilisant la méthode H-FWER. En outre, les diagnostics de pente et de convexité ont été étudiés simultanément en utilisant la méthode H-FWER pour la convexité en contrôlant le taux d'erreur de type I global à $\alpha = 0,05$ et à $\alpha = 0,10$, tandis que la pente a été évaluée à $\delta = 0,5$ et $\delta = 0,10$. Les noyaux quantique et TH ont tous deux été utilisés pour les deux tests (voir la section 4.1)

Table pour H-FWER pour pics multiples											
n	C = 0,05		C = (0,10, 0,05)		(S, C) = (0,25, 0,5)		C = 0,10		(S, C) = (0,10, 0,10)		TH
	Quantique	TH	Quantique	TH	Quantique	TH	Quantique	TH	Quantique	TH	
120	0,006	0,002	0,006	0,002	0,005	0,002	0,076	0,047	0,070	0,044	0,076
144	0,009	0,002	0,011	0,002	0,008	0,003	0,087	0,053	0,090	0,051	0,086
180	0,019	0,005	0,020	0,006	0,017	0,005	0,107	0,062	0,097	0,059	0,093
288	0,031	0,009	0,026	0,008	0,025	0,008	0,117	0,069	0,116	0,069	0,112
360	0,042	0,012	0,045	0,019	0,035	0,015	0,140	0,087	0,133	0,084	0,129
	0,087										0,087

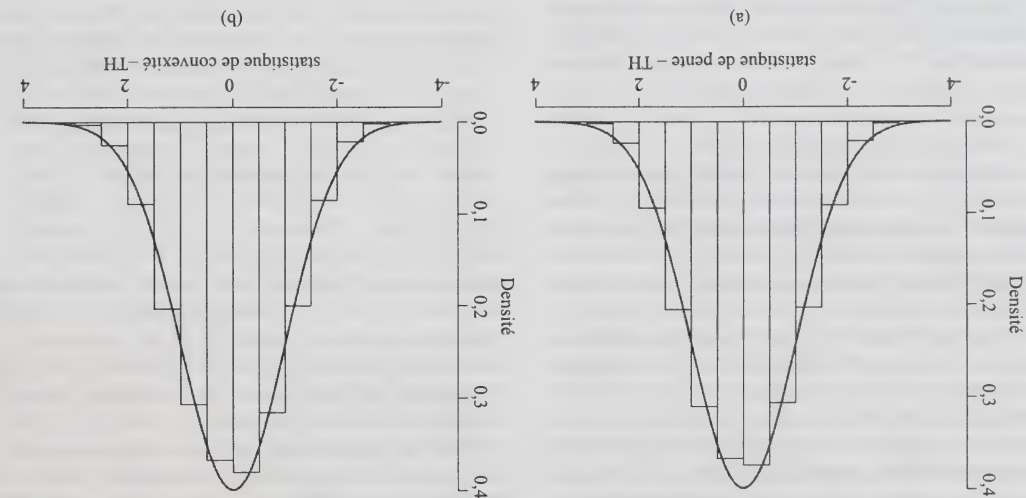


Figure 1 Histogramme de la distribution de $-\sqrt{n}\psi_{h_n}^{(I)}$ (a) et de $-\sqrt{n}\psi_{h_n}^{(I)}$ (b) sous une hypothèse nulle de bruit blanc gaussien en utilisant le noyau TH. La taille d'échantillon est $n = 360$ avec 10 000 répétitions.

Tableau 1 Résultats de la simulation de la taille pour le diagnostic d'un pic unique. Ici $\mu = \beta = \pi/6$ et 10 000 répétitions sont utilisées. Les diagnostics de pente et de convexité ont été étudiés séparément pour les noyaux quartique et TH

n	Noyau quartique				Noyau TH				Convexité			
	Moyenne	Écart-type	Niveau δ	Moyenne	Écart-type	Niveau δ	Moyenne	Écart-type	Moyenne	Écart-type	Niveau α	Moyenne
120	0,003	0,903	0,007	-0,011	0,903	0,008	-0,065	0,852	0,032	0,025	0,888	0,018
144	-0,004	0,920	0,014	-0,011	0,927	0,015	-0,077	0,882	0,042	0,006	0,892	0,025
180	-0,003	0,942	0,022	0,002	0,920	0,017	-0,071	0,892	0,043	0,005	0,902	0,028
288	0,003	0,954	0,027	-0,002	0,950	0,025	-0,072	0,921	0,051	-0,006	0,926	0,033
360	0,003	0,962	0,032	-0,009	0,954	0,031	-0,056	0,922	0,051	0,006	0,951	0,040

Taille pour un pic unique $\mu = \beta = \pi/6$

Tableau 2 Résultats de la simulation de la puissance pour le diagnostic d'un pic unique. Ici $\mu = \beta = \pi/6$ et 10 000 répétitions sont utilisées. L'hypothèse alternative est donnée par le modèle $AR(2)$ défini par (11). Les diagnostics de pente et de convexité ont été étudiés simultanément pour les noyaux quartique et TH en utilisant $\delta = \alpha = 0,05$ pour les deux tests (voir la section 4.1)

n	Noyau quartique				Noyau TH			
	$p = 0,85$	$p = 0,90$	$p = 0,95$	$p = 0,85$	$p = 0,90$	$p = 0,95$	$p = 0,85$	$p = 0,95$
120	0,227	0,338	0,758	0,147	0,335	0,670	0,335	0,670
144	0,287	0,332	0,856	0,208	0,431	0,799	0,431	0,799
180	0,354	0,434	0,923	0,272	0,567	0,901	0,567	0,901
288	0,447	0,755	0,949	0,372	0,706	0,950	0,706	0,950
360	0,601	0,872	0,937	0,537	0,859	0,948	0,859	0,948

Puissance pour un pic unique $\mu = \beta = \pi/6 - (\delta, \alpha) = (0,05, 0,05)$

produisent une distribution empirique des diagnostics normalisés, $\psi_A(I)$ et $\psi_B(I)$, dont les histogrammes sont présentés à la figure 1. À partir de maintenant, représentons par δ et α les niveaux associés aux tests de pente et de convexité, respectivement. Notons que, dans ce cas, nous définissons le niveau comme étant la probabilité de rejeter $H_0^p(I) = 1, 2$ quand H_0^p est vraie. Même si, en pratique, dans le cas de l'hypothèse de pente, nous soustirons ne pas aboutir au rejet, nous suivons la définition stricte du niveau et supposons (pour les besoins de cette simulation) que l'hypothèse nulle H_0^p pour la pente est vérifiée. De même, l'hypothèse nulle pour la convexité est H_0^c . Les hypothèses de pente et de convexité sont évaluées indépendamment. Le tableau 1 résume les résultats en utilisant les deux niveaux décrits à la section 2, pour diverses tailles d'échantillons ; α indiqués s'appliquent au seuil nominal de 5 %. En outre, d'autres choix de μ et β (non présentés ici) ont donné des résultats semblables. Comme l'illustre la présente étude, dans les échantillons de petite taille, nous observons une asymétrie de la distribution qui semble être due à la corrélation entre $\theta_A(I)$ et $\theta_B(I^2)$. En outre, nous que dans le cas du test de convexité, la taille est plus grande pour le noyau quartique que pour le noyau TH.

Ensuite, considérons la puissance empirique de notre diagnostic de la présence d'un seul pic. Dans ces conditions, nous évaluons la puissance basée sur un test conjoint de la pente et de la convexité. Plus précisément, nous soustirons ne pas rejeter $H_0^{(1)}$, tout en rejetant simultanément $H_0^{(2)}$, au niveau $\delta = \alpha = 0,05$, et donc identifier correctement les pics spectraux. Puisque notre hypothèse nulle composite est qu'il n'existe pas de pic, l'hypothèse alternative inclut les processus tels que le processus $AR(2)$ donné par

$$(1 - 2\rho \cos \omega B + \rho^2 B^2)X'_t = \varepsilon_t \quad (11)$$

avec la variance de bruit blanc σ_ε^2 , associée à une fréquence fixe donnée $\omega \in [0, \pi]$. Le spectre associé au processus (11) est donné par $f(\lambda) = \sigma_\varepsilon^2 |1 - 2\rho \cos \omega e^{-i\lambda} + \rho^2 e^{-2i\lambda}|^{-2}$, qui est maximisée à $\lambda_0 = \cos^{-1}(\cos \omega (1 + \rho^2/2\rho))$. Par conséquent, nous pouvons explorer la puissance d'une procédure de dépistage de pic en procédant à une simulation selon (11) avec divers choix de ρ , ω et α . Le tableau 2 donne les résultats de 10 000 simulations, pour diverses tailles d'échantillons, d'après le modèle cyclique $AR(2)$ donné en (11) avec un pic à $\mu = \pi/6$ et la largeur de bande fixée à $\beta = \pi/6$. La force du pic est paramétrisée au moyen de ρ , que nous faisons varier de 0,85 à 0,95 ; il est clair que $H_0^{(1)}$ et $H_0^{(2)}$ sont toutes deux vraies pour ce modèle. Autrement dit, il existe des pics spectraux, de

différentes hauteurs, à $\lambda = \pi/6$. Ce modèle AR a été choisi parce qu'il donne une paramétrisation commode de la localisation et de la forme du pic spectral. De plus, ce choix de β est compatible avec les conditions de désaisonnalisation, car il fournit la largeur de fenêtre maximale tout en évitant le chevauchement des pics spectraux. Comme prévu, la puissance de notre diagnostic augmente avec la taille de l'échantillon et la pointinité, variant de 0,227 (noyau quartique) dans les petits échantillons possédant un faible pic spectral à $\approx 0,95$ (noyau TH) dans les échantillons plus grands ayant un pic spectral plus prononcé (voir le tableau 2). Notons que, dans cette procédure, la valeur de la variance d'innovation est fixée à 1, mais qu'elle est négligeable en raison de la normalisation du diagnostic. En résumé, les noyaux quartique et TH possèdent tous deux des propriétés de taille et de puissance convenables. Généralement, le noyau quartique semble avoir une taille et une puissance supérieures, si bien qu'on le préférerait pour des spectres de cette forme (notons que la puissance plus faible du noyau TH est due en partie au fait que sa taille est trop petite). En outre, il semble que les valeurs plus faibles de β (résultats non présentés) requièrent une plus grande taille d'échantillon ; un β plus petit correspond à des « conditions d'observation » plus fines du pic spectral, qui nécessiterait une plus grande quantité de données pour traiter la résolution.

Bien que le scénario du test de la présence de pics individuels constitue le fondement de notre cadre de test conjoint, comme nous l'avons mentionné, ce cadre représente une méthodologie importante pour les applications aux statistiques fédérales. L'application vraiment importante est l'exploration de la saisonnalité résiduelle en vue de déterminer si la désaisonnalisation est efficace. Il est donc d'approche de tests multiples se comporte dans la simulation. Par conséquent, pour étudier la taille et la puissance associées à notre test conjoint, nous avons simulé 10 000 répétitions à partir d'un processus de bruit blanc gaussien et d'un modèle $AR(25)$ obtenu par un ajustement à la série sur l'emploi courant (hommes occupés, de 16 à 19 ans). L'objectif de notre étude de puissance était de construire un processus $AR(p)$ (à cause de la facilité de son utilisation dans les simulations et de ses propriétés théoriques désirables en tant qu'estimateur spectral paramétrique – voir Parzen 1983) avec des pics spectraux (stationnaires) qui sont réalistes, ou qui s'approchent de ce qu'on pourrait trouver en pratique. Donc, nous obtenons notre modèle $AR(25)$ – ajusté par la méthode du maximum de vraisemblance en utilisant le critère AIC – qui possède la même dynamique saisonnière (comportement spectral local) que la série sur l'emploi courant (CES pour *Current Employment Series*).

statistiques de test de convexité (voir la discussion qui suit le théorème 1 à l'annexe). Naturellement, nous exécutons aussi cinq tests distincts de la pente à chaque fréquence saisonnière, où il faut que nous n'abouissions à un rejet dans aucun cas pour pouvoir poursuivre.

Enfin, nous notons qu'en pratique, une désaisonnalisation est rarement rejetée sur la base d'une masse spectrale significative à la cinquième fréquence saisonnière de $5\pi/6$ (Findley 2006). Cela tient en partie à la difficulté d'attribuer une interprétation à cette fréquence. Par conséquent, le praticien pourrait être plus intéressé par un « test à quatre pics » qui est axé sur les quatre premières fréquences saisonnières ; ce test s'obtient par une modification évidente de la procédure H-FWER décrite plus haut.

3.4 Extension à des données non stationnaires

La méthodologie que nous venons d'exposer s'appuie sur l'hypothèse que les données sont un échantillon d'un processus stationnaire. Cependant, dans le contexte de la désaisonnalisation, les données désaisonnalisées sont habituellement intégrées une fois ou deux fois. Dans ce cas, nous différencierait les données désaisonnalisées une ou deux fois avant d'appliquer les diagnostics. Or, les opérateurs de différentiation $1 - B$ et $(1 - B)^2$ appliqués sont essentiellement des filtres passe-haut, qui en principe atténuent les pics spectraux résiduels proches de la fréquence zéro (en particulier, la première fréquence saisonnière à $\pi/6$). Donc, il serait peut-être souhaitable d'appliquer le diagnostic au pseudo-spectre, ce qui peut se faire si le support du moyen spectral.

Supposons que les données observées sont maintenant X_{1-d}, \dots, X_n pour d l'ordre de différenciation de la tendance (donc habituellement $d = 1$ ou 2). Lorsque les données observées sont différenciées, nous obtenons l'échantillon X , qui est strictement stationnaire. Le pseudo-densité spectrale du processus $\{X_t\}$ est $g(\lambda) = f(\lambda) |1 - e^{-i\lambda}|^{-2d}$, où f est le spectre de $\{X_t\}$. Ce pseudo-spectre pourrait être estimé au moyen de $\hat{g}(\lambda) = I(\lambda) |1 - e^{-i\lambda}|^{-2d}$, où I est le périodogramme de X comme auparavant ; il s'agit de l'approche de recoloration de Nerlove (1964). Alors, $\theta_d(g)$ est bien défini à condition que $N(\lambda) |1 - e^{-i\lambda}|^{-2d}$ soit une fonction intégrable ; essentiellement, nous devons nous assurer que la fréquence zéro est exclue du support du moyen λ . Puisqu'en pratique λ est centré autour des fréquences saisonnières, nous pouvons facilement trouver le moyen d'appliquer cette condition. L'estimateur correspondant est alors

$$\theta_d(g) = \theta_d(I).$$

où $b(\lambda) = |1 - e^{-i\lambda}|^{-2d}$. L'estimateur est bien défini si λb est intégrable ; en outre, les propriétés asymptotiques

discutées à l'annexe pour le cas stationnaire s'étendent à ce cas également, à condition que λb soit borné. Cette extension pourrait paraître plus séduisante à certains chercheurs. Cependant, son prix est que la transformée de Fourier inverse de λb doit être déterminé, ce qui requiert un travail mathématique supplémentaire. Dans les études par simulation et les illustrations des données de la section 4, nous différencions les données désaisonnalisées, mais nous n'appliquons pas le facteur de correction b dans le moyen.

4. Études empiriques

Ayant développé les aspects théoriques du diagnostic spectral, nous nous tournons maintenant vers sa performance en pratique. Pour commencer, nous présentons certains résultats obtenus par simulation, qui donnent une idée des propriétés de taille et de puissance de la statistique de test dans les échantillons finis. Puis, nous étudions empiriquement la taille et la puissance en appliquant les diagnostics spectraux à une suite de 130 séries chronologiques (65 séries du U.S. Census Bureau et 65 séries de l'OCDE) ; nous considérons les séries originales ainsi que les séries désaisonnalisées et nous comparons les résultats à ceux du diagnostic de signification visuelle et aux diagnostics de contrôle de la qualité M7 et M8 de X-12-ARIMA (U.S. Census Bureau 2002). D'autres études empiriques sont décrites dans Evans, Holan et McElroy (2006).

4.1 Étude par simulation

Afin d'évaluer la performance de nos diagnostics, nous avons exécuté plusieurs simulations. Le premier ensemble de simulations a pour but d'examiner la taille (niveau) pour le diagnostic de la présence d'un pic unique. Pour cette simulation, nous considérons les diagnostics de pente et de convexité séparément. Même si, en pratique, quand nous considérons le diagnostic de pente, nous souhaitons ne pas rejeter l'hypothèse nulle $H_0^{(1)}$, ici nous voulons étudier empiriquement les propriétés distributionnelles et nous imposons donc la définition habituelle de taille pour cette étude. Donc, nous simulons un bruit blanc gaussien qui satisfait les hypothèses du théorème 1, ainsi que $\psi_d(I) = \psi_d(I) = 0$, de sorte que $H_0^{(1)}$ et $H_0^{(2)}$ sont vraies. Naturellement, il existe de nombreux processus pour lesquels $H_0^{(1)}$ et $H_0^{(2)}$ sont vraies simultanément – par exemple, tout processus dont la densité spectrale est localement uniforme ; suffit de considérer le bruit blanc. Pour une (grande) taille cependant, en raison de considérations asymptotiques, il suffit de considérer le bruit blanc. Pour une (grande) taille d'échantillon de $n = 360$, en utilisant le moyen TH avec $\pi = \pi/6$ et $\beta = \pi/6$ (ce qui correspond à un moyen centré sur l'intervalle $[0, \pi/6]$), 10 000 répétitions

Degré de non-normalité dans les petits échantillons. D'après l'histogramme de la distribution simulée sous une hypothèse nulle de bruit blanc gaussien avec $n = 360$ et 10 000 répétitions (figure 1), la concordance avec la loi normale est proche, sauf aux extrémités dans les queues. À la section 4, nous examinons ce comportement plus en détail au moyen d'études par simulation.

3.2 Applications au test de la présence d'un seul pic

Nous considérons maintenant l'application au test de la présence d'un pic. Rappelons que nous avons une hypothèse nulle initiale $H_0^{(1)}$ qui ne doit pas être rejetée pour pouvoir poursuivre. Cela peut s'interpréter comme signifiant qu'il n'existe pas de preuve suffisamment convaincante pour conclure que la dérivée première (pente) de la densité spectrale diffère de manière significative de zéro. Or, nous savons que $-\sqrt{n}\psi_{A_p,n}(I)$ est asymptotiquement de loi $N(0, 1)$ sous $H_0^{(1)}$, ainsi que les hypothèses discutées à l'annexe. Si nous supposons en outre qu'une valeur suffisamment petite x est obtenue pour la statistique de test, nous ne pourrions rejeter $H_0^{(1)}$ avec aucun degré de confiance. Dans ce cas, nous pouvons considérer l'hypothèse $H_0^{(2)}$, que nous cherchons à rejeter, ce que nous testons au moyen de $-\sqrt{n}\psi_{A_p,n}(I)$. Bien que $-\sqrt{n}\psi_{A_p,n}(I)$ et $\sqrt{n}\psi_{A_p,n}(I)$ soient asymptotiquement corrélées (voir le théorème 1 de l'annexe), nous considérons les tests de la pente et de la convexité comme s'ils étaient exécutés séparément. (Cette corrélation peut être estimée et utilisée pour déterminer la distribution du diagnostic de convexité conditionnellement au diagnostic de pente; cependant, l'interprétation des valeurs p devient confuse. Pour simplifier, nous traitons les tests séparément, un à la fois, et nous ne tenons pas compte explicitement de la corrélation.) Notre procédure de test s'exécute alors de la façon suivante :

1. Faire le test bilatéral de $H_0^{(1)}$ en utilisant $-\sqrt{n}\psi_{A_p,n}(I)$.
2. Poser que p est la valeur p associée avec la première valeur de la statistique de test $x = -\sqrt{n}\psi_{A_p,n}(I)$, avec x et p reliées par $p = 2\Phi(-|x|)$.
3. Si $p > 0,05$ (ou un autre seuil de tolérance pré-établi), poursuivre; sinon, conclure qu'aucun pic n'est présent.
4. Exécuter le test unilatéral inférieur de $H_0^{(2)}$ en utilisant $\sqrt{n}\psi_{A_p,n}(I)$.
5. Rejeter $H_0^{(2)}$ et conclure qu'il existe un pic si $\sqrt{n}\psi_{A_p,n}(I) > \Phi^{-1}(\alpha)$, où α est le niveau du test de convexité.

3.3 Test conjoint de la présence de plusieurs pics :

Application à la désaisonnalisation

Considérons maintenant la situation où nous voulons tester simultanément la présence de plusieurs pics spectraux. Il est clair que nous pourrions concevoir un moyen comportant plusieurs nœuds, un à chaque pic, mais il serait simplement la somme de plusieurs diagnostics de pic spectral individuels. L'inconvénient serait qu'un pic spectral significatif à un nœud pourrait annuler un creux spectral significatif ailleurs. Par conséquent, nous préférons un test permettant d'examiner un ensemble de diagnostics spectraux dans un paradigme de tests multiples.

Par exemple, considérons le contexte de la détection de la présence de pics spectraux dans des données désaisonnalisées. Six pics saisonniers présentent un intérêt, mais nous devons nous limiter à cinq d'entre eux à cause de problèmes d'allias, c'est-à-dire de repliement du spectre (le pic de fréquence π ne peut pas être identifié). Si un ou plusieurs pics spectraux sont significatifs, nous devons rejeter notre procédure de désaisonnalisation (puisqu'elle n'a pas réussi à éliminer tous les pics); par conséquent, nous sommes dans des conditions de tests multiples et nous utilisons une méthode permettant de contrôler le taux d'erreur de type I global (FWER pour *familywise error rate*) proposé par Hochberg (1988) et décrit dans Benjamini et Hochberg (page 294, 1995). En nous limitant à la question de la convexité, nous avons les hypothèses nulles $H_0^{(2)}$ pour chacune des cinq fréquences saisonnières. Dans nos conditions, la procédure de Hochberg (1988) consiste à calculer les valeurs p pour le test de convexité à chacune de ces cinq fréquences saisonnières et à les classer selon $p^{(1)} \leq p^{(2)} \leq p^{(3)} \leq p^{(4)} \leq p^{(5)}$ avec les hypothèses nulles correspondantes désignées par $H_0^{(i)}$. Pour un taux d'erreur de type I global spécifié de niveau α (par exemple $\alpha = 0,05$), soit k la valeur de i la plus grande pour laquelle $p^{(i)} \leq i/(6 - i)$; alors, nous rejetons toutes les $H_0^{(i)}$ pour $i \leq k$.

En utilisant cette procédure, nous devons commettre des erreurs de type I – c'est-à-dire déterminer la présence d'au moins une fréquence saisonnière ayant une convexité négative alors qu'il n'en existe aucune – une proportion de fois environ égale à α (si nous nous limitons à examiner $H_0^{(2)}$, l'hypothèse de convexité). L'avantage de l'approche FWER) est qu'elle améliore spectaculairement la puissance statistique comparativement aux autres méthodes. La validité de cette méthode requiert l'indépendance des statistiques de test prises en considération et, pour cette raison, nous prenons cinq moyaux A_1, \dots, A_5 – centres sur les fréquences saisonnières $\pi/6, \dots, 5\pi/6$ respectivement – qui ont un support disjoint. Alors, le théorème 1 peut être généralisé pour obtenir l'indépendance asymptotique de cinq

propriétés asymptotiques. À la section 3.2, nous discutons de l'application à la détection individuelle des pics et à la section 3.3, nous donnons une extension à leur détection conjointe, qui facilite une application importante dans la désaisonnalisation. À la section 3.4, nous discutons des extensions à des données non stationnaires avec tendance.

3.1. Estimateurs de la pente et de la convexité

Nous commençons par noter que la forme quadratique (pour toute fonction intégrable g) est

$$\frac{1}{I} X' \Sigma(g) X = \frac{2\pi}{1} \int_{-\pi}^{\pi} g(\lambda) I(\lambda) d\lambda,$$

où I désigne le périodogramme. Bien que ce dernier soit habituellement défini par les fréquences de Fourier $(2\pi/n; j = 1, \dots, \lfloor n/2 \rfloor)$, nous le définissons comme une bande continue de fréquences de la façon suivante

$$I(\lambda) = \frac{1}{2} \left| \sum_{j=1}^n X_j' e^{-i\lambda j} \right|^2 = \sum_{j=1}^{n-1} R(h) e^{-i\lambda h}, \lambda \in [-\pi, \pi] \quad (9)$$

avec $R(h)$ égal à la fonction d'autocovariance (non centrée) d'échantillon. Cela donne un moyen élégant de passer du domaine temporel au domaine fréquentiel qui est bien décrit dans la littérature sur les séries chronologiques (voir Taniguchi et Kakizawa 2000). En outre, ce genre d'intégrales du périodogramme sont généralement convergentes, c'est-à-dire que $\theta_g^g(I) \xrightarrow{P-X} \theta_g^g(f)$ quand $n \rightarrow \infty$, sous les conditions faibles discutées plus bas (notons que la non-convergence du périodogramme est résolue par l'agrégation spectrale contre la fonction g , comme il est montré en annexe). Par conséquent, nous obtenons des estimations statistiques des mesures de la pente et de la convexité f en utilisant une approche de substitution, c'est-à-dire en remplaçant simplement f par I dans $\theta_{A_{B,n}}^g$. En particulier,

$$\theta_{A_{B,n}}^g(f) = -\theta_{A_{B,n}}^g(I) = -\frac{1}{I} X' \Sigma(A_{B,n}) X, \text{ et} \quad (10) \quad \theta_{A_{B,n}}^g(f) = \theta_{A_{B,n}}^g(I) = \frac{1}{I} X' \Sigma(A_{B,n}) X.$$

Cette définition comprend l'utilisation de (5), qui explique le signe négatif de la mesure de la pente. Afin de calculer l'estimation, nous utilisons la représentation dans le domaine temporel (exprimée sous une forme quadratique). Cette représentation est commode parce que nous devons uniquement déterminer une longueur appropriée pour les séquences $\gamma_{A_{B,n}}(h)$ et $\Sigma(A_{B,n})(h)$, former les matrices de Toeplitz $\Sigma(A_{B,n})$ et $\Sigma(A_{B,n})$, puis calculer les formes quadratiques. Notons que les transformations de Fourier inverses de $A_{B,n}$ et $A_{B,n}$ ne doivent être déterminées qu'une seule fois (voir la section 2.3 pour des exemples explicites) et peuvent être produites d'avance.

Afin de calculer la représentation dans le domaine chronologiques différentes puis appliquées de manière répétée à de nombreuses séries temporel des mesures de la pente et de la convexité données par (10), nous utilisons (8) ; par exemple voir les formules dans les exemples 1 et 2. Évidemment, il résultera en général que $\Sigma(A_{B,n})$ et $\Sigma(A_{B,n})$ sont complexes. Cependant, même si $\Sigma(g)$ (où g peut être $A_{B,n}$ ou $A_{B,n}$) est une matrice de Toeplitz complexe, $X' \Sigma(g) X$ sera toujours réel. Partant de (8), il est facile de voir que $\Sigma(g) = M + iN$ où M est réel, symétrique et Toeplitz, et que N est réelle, antisymétrique et Toeplitz. D'où $X'NX = 0$ pour tout vecteur X , de sorte que $X' \Sigma(g) X = X'MX$. Par conséquent, pour calculer les mesures statistiques de la pente et de la convexité, nous pouvons prendre la partie réelle de $\gamma^g(h)$ dans (8).

Ces estimations statistiques sont non seulement convergentes, mais aussi asymptotiquement normales sous certaines conditions supplémentaires (discutées à l'annexe). Cependant, pour construire une normalisation appropriée, il sera nécessaire d'estimer leur variation. La variance asymptotique de $\theta_g^g(I)$ est $\theta_g^g(f^2)$ (si g est appuyée sur $[0, \pi]$), qui peut être estimée de manière convergente par la voie de $\theta_{g^g}(I^2)/2$. (Le facteur de 2 est requis, puisque l'intégrale de I^2 tend vers l'intégrale correspondante de I^2 tend vers l'intégrale correspondante de $2f^2$ – voir Chiu (1988)). Nous pouvons représenter cela dans le domaine temporel de la façon suivante. Soit $R = \{R(1 - n), \dots, R(0), \dots, R(n - 1)\}^T$ un vecteur de dimension $2n - 1$ des autocovariances d'échantillon et soit $\Sigma(g^2)$ de dimension $2n - 1$ dans la formule suivante : $R' \Sigma(g^2) R / 2 = \theta_{g^g}(I^2) / 2$. Cette relation peut être vérifiée facilement en utilisant (9). Donc, nous normalisons $\theta_g^g(I)$ par la racine carrée de $\theta_{g^g}(I^2) / 2$. D'où une mesure statistique normalisée de la pente et de la convexité sont données par

$$-\psi_{A_{B,n}}(I) = -\frac{\theta_{A_{B,n}}(I)}{\theta_{A_{B,n}}(I^2)/2} = \frac{1}{I} X' \Sigma(A_{B,n}) X / \sqrt{R' \Sigma(A_{B,n}^2) R / 2},$$

et

$$\psi_{A_{B,n}}(I) = \frac{\theta_{A_{B,n}}(I)}{\theta_{A_{B,n}}(I^2)/2} = \frac{1}{I} X' \Sigma(A_{B,n}) X / \sqrt{R' \Sigma(A_{B,n}^2) R / 2},$$

où les dimensions des matrices Σ sont soit n soit $2n - 1$ comme il est approprié. Les propriétés asymptotiques de $\psi_{A_{B,n}}(I)$ et de $\psi_{A_{B,n}}(I)$ sont discutées à l'annexe. En résumé, $-\psi_{A_{B,n}}(I)$ et $\psi_{A_{B,n}}(I)$ sont toutes deux marginalement asymptotiquement de loi $N(0, 1)$ sous $H_0^{(1)}$ et $H_0^{(2)}$, respectivement, et sous les hypothèses discutées à l'annexe. Les simulations indiquent que la normalisation de la variance ne converge que lentement et que sa corrélation avec le numérateur cause un certain

$$\gamma_A(h) = \frac{\pi^5}{15!} \left(\pi^2 \sin \pi h + 3\pi \cos \pi h - \frac{h^2}{3 \sin \pi h} \right) + \frac{\pi^5}{15} \left(\pi^2 \sin \pi h + 3\pi \cos \pi h - \frac{h^2}{3 \sin \pi h} \right),$$

$$\gamma_{A^2}(h) = \frac{\pi^9}{225} \left(-2\pi^4 \sin \pi h - \frac{h^3}{18\pi^3 \cos \pi h} \right) + \frac{\pi^9}{225} \left(2\pi^4 \sin \pi h + \frac{h^3}{18\pi^3 \cos \pi h} \right) + \frac{4\pi^{11}}{225} \left(\pi^4 \sin \pi h + 6\pi^3 \cos \pi h - \frac{h^2}{6\pi^2 \sin \pi h} - \frac{h^2}{54 \sin \pi h} \right),$$

auxquels nous appliquons (8) et obtenons

$$\gamma_{A^3}(h) = \frac{\beta}{30!} \exp\{i\pi h\} \left(\frac{\sin k}{3 \cos k} + \frac{\pi^2}{3 \sin k} - \frac{\pi^4}{3 \sin k} \right), \quad \gamma_{A^3}(h) = \frac{\beta^2}{30} \exp\{i\pi h\} \left(\frac{\sin k}{3 \cos k} + \frac{\pi^2}{3 \sin k} - \frac{\pi^4}{3 \sin k} \right), \quad \gamma_{A^3}(h) = \frac{\beta^3}{1800\pi} \exp\{i\pi h\} \left(\frac{2 \sin k}{18 \cos k} + \frac{\pi^3}{180 \cos k} - \frac{\pi^5}{180 \sin k} - \frac{\pi^7}{180 \sin k} \right), \quad \text{et}$$

$$\gamma_{A^3}(h) = \frac{1}{800} \beta^5 \exp\{i\pi h\} \left(\frac{\sin k}{6 \cos k} + \frac{\pi^5}{54 \cos k} - \frac{\pi^7}{54 \sin k} \right),$$

où $k = h\beta/2$. Notons que $\gamma_{A^3}(0) = 240\pi/(7\beta^3)$ et $\gamma_{A^3}(0) = 360/(\beta^3\pi)$ découlent de l'application de la règle de L'Hôpital. Ces formules nous permettent de construire les matrices de Toeplitz appropriées pour le diagnostic (comme il est exposé plus bas, à la section 3.1, il suffit de considérer la partie réelle de ces séquences).

Exemple 2 : Noyau TH

Nous pouvons obtenir une forme semblable au noyau quartique en utilisant une fonction cosinus. Le choix qui suit satisfait toutes les conditions énoncées sur un noyau :

$$\begin{aligned} \lambda(h) &= \frac{2\pi}{1} (1 + \cos \lambda), \\ \lambda(h) &= \frac{1}{2\pi} (-\sin \lambda), \text{ et} \\ \lambda(h) &= \frac{1}{2\pi} (-\cos \lambda). \end{aligned}$$

Cette fonction est identique à la fenêtre des décalages de Tukey-Hanning, quoique ici, nous l'appliquons comme une

3. Méthodologie statistique

où $k = h\beta/2$. Notons que $\gamma_{A^3}(0) = \pi/\beta^3$ et $\gamma_{A^3}(0) = 4\pi^3/\beta^5$ découlent de l'application de la règle de L'Hôpital (en utilisant la convention que $\sin(0)/0 = 1$). Ces formules nous permettent de construire les matrices de Toeplitz appropriées pour le diagnostic (de nouveau, comme nous l'exposons plus bas à la section 3.1, il suffit de considérer la partie réelle de ces séquences).

$$\begin{aligned} \gamma_{A^3}(h) &= \frac{2\beta^3}{\pi} \exp\{i\pi h\} \left(\frac{2 \sin k}{2 \sin k} - \frac{k}{\sin(k+2\pi)} + \frac{\sin(k-2\pi)}{k-2\pi} \right), \text{ et} \\ \gamma_{A^3}(h) &= -\frac{\beta^2}{\pi} \exp\{i\pi h\} \left(\frac{\sin(k+\pi)}{\sin(k-\pi)} + \frac{k+\pi}{\sin(k-\pi)} \right), \\ \gamma_{A^3}(h) &= \frac{2\beta^3}{i} \exp\{i\pi h\} \left(\frac{\sin(k+\pi)}{\sin(k-\pi)} - \frac{k+\pi}{\sin(k-\pi)} \right), \end{aligned}$$

L'application de (8) donne alors

$$\begin{aligned} \gamma_A(h) &= \frac{4\pi^2}{i} \left(\frac{\sin \pi(h+1)}{\sin \pi(h-1)} - \frac{h+1}{\sin \pi(h-1)} \right), \\ \gamma_A(h) &= -\frac{1}{4\pi^2} \left(\frac{h+1}{\sin \pi(h+1)} + \frac{h-1}{\sin \pi(h-1)} \right), \\ \gamma_{A^2}(h) &= \frac{1}{2 \sin \pi h} - \frac{h}{2 \sin \pi(h+2)} - \frac{h}{2 \sin \pi(h-2)}, \text{ et} \\ \gamma_{A^2}(h) &= \frac{1}{16\pi^3} \left(\frac{h}{2 \sin \pi h} + \frac{h}{\sin \pi(h+2)} + \frac{h}{\sin \pi(h-2)} \right). \end{aligned}$$

la convexité (et leurs carrés), nous obtenons transformée de Fourier inverse des noyaux de la pente et de l'exposé, nous la désignerons par noyau TH. En prenant la fenêtre spectrale (voir Priestley 1981). Dans la suite de

$$\hat{A}_{2,\mu}^{\beta,\mu}(\lambda) = \frac{64\pi^5}{\beta} \hat{A}_{2,\mu}^{\beta} \left(\frac{\beta}{2\pi} (\lambda - \mu) \right).$$

Enfin, nous notons d'après (4) que nous pouvons réécrire $\theta_A(f)$ sous la forme

$$\theta_A(f) = \sum_{h=-\infty}^{\infty} \gamma_A(h) \gamma_f(h). \quad (7)$$

Donc, il pourrait être avantageux de déterminer la séquence $\gamma_A(h)$ à partir du noyau A . En prenant la transformée de Fourier inverse des noyaux susmentionnés de la pente et de la convexité, nous pouvons construire $\Sigma(A_{\beta,\mu}^{\beta,\mu})$, $\Sigma(A_{\beta,\mu}^{\beta,\mu})$ et $\Sigma(A_{2,\mu}^{\beta,\mu})$ comme il suit :

$$\gamma_{A_{\beta,\mu}^{\beta,\mu}}(h) = \frac{\beta}{2\pi} \exp\{i h \mu\} \gamma_A(h \beta / 2\pi),$$

$$\gamma_{A_{\beta,\mu}^{\beta,\mu}}(h) = \frac{\beta^2}{4\pi^2} \exp\{i h \mu\} \gamma_A(h \beta / 2\pi),$$

$$\gamma_{A_{2,\mu}^{\beta,\mu}}(h) = \frac{\beta^3}{8\pi^3} \exp\{i h \mu\} \gamma_A(h \beta / 2\pi),$$

$$\gamma_{A_{2,\mu}^{\beta,\mu}}(h) = \frac{\beta^5}{32\pi^5} \exp\{i h \mu\} \gamma_A(h \beta / 2\pi). \quad (8)$$

Donc, si nous possédons l'information relative au domaine temporel $\gamma_f(h)$ pour le processus $\{X_t\}$, nous pouvons calculer les mesures de la pente et de la convexité en utilisant (7), sachant la séquence des transformées de Fourier inverses des noyaux appropriés. Puisque $\gamma_f(h)$ est une séquence symétrique, nous ne devons considérer que la partie réelle de $\gamma_A(h)$ s'il se fait que sa valeur est complexe.

2.2 Creux et pics

Les mesures agrégées de la pente et de la convexité spectrales décrites antérieurement représentent les composantes élémentaires des déterminants de la géométrie spectrale locale. Notre objectif général est de déterminer si un intervalle donné du spectre correspond à un pic ou à un creux (ou est monotone). Dans la géométrie de deuxième ordre de l'analyse infinitésimale, un maximum local possède la propriété déterminante que sa dérivée première est nulle et sa dérivée seconde, strictement négative. Il faut donc manifestement examiner, séquentiellement, une mesure de pente et une mesure de convexité définies sur la même bande de fréquences.

Afin de vérifier la présence d'un pic, l'approche séquentielle peut être considérée comme équivalente à formuler des énoncés inférentiels au sujet de $\theta_{A_{\beta,\mu}^{\beta,\mu}}(f)$ et $\theta_{A_{2,\mu}^{\beta,\mu}}(f)$. Soulignons qu'en formulant ces énoncés inférentiels, nous choisissons la valeur de μ d'avance, en fonction de l'endroit dans le spectre où nous souhaitons détecter un pic (ou un creux) ; β est choisi en fonction des fréquences que

nous désirons exclure, décision qui dépend de la mesure dans laquelle nous souhaitons que notre point de vue du spectre soit local. Puis, nous disons que μ est un pic β -agrégé (par rapport à A) du spectre si

$$\theta_A(f) = 0 \quad \text{et} \quad \theta_A(f) < 0.$$

L'aspect séquentiel découle de l'idée que, en général, nous commençons par déterminer si $\theta_A(f) = 0$, puis que nous déterminons la convexité ; cela devientra plus évident quand nous examinerons les tests statistiques à la section 3.2. De manière comparable, nous définissons un creux β -agrégé quand $\theta_A(f) > 0$. En ce qui concerne le test d'hypothèse pour un pic, nous avons

$$\begin{aligned} H_{(1)}^0 : \theta_A(f) &= 0 & \text{vs} & & H_{(1)}^a : \theta_A(f) \neq 0 \\ H_{(2)}^0 : \theta_A(f) &= 0 & \text{vs} & & H_{(2)}^a : \theta_A(f) > 0. \end{aligned}$$

L'aspect inhabituel de ce test d'hypothèse est que nous voulons d'abord ne pas rejeter $H_{(1)}^0$, puis, conditionnellement à ce test, nous voulons rejeter $H_{(2)}^0$ en faveur de l'alternative $H_{(2)}^a$.

2.3 Exemples de noyaux

Il existe une foule de noyaux qui satisfont les conditions (i) à (iv) ; il nous suffit de les emprunter à la littérature sur l'estimation non paramétrique de la densité. Par exemple, dans Priestley (1981) conveniement, tandis que ceux de Bartlett et Danielli ne sont pas appropriés, puisque (iv) n'est pas vérifiée. En général, il suffit d'utiliser (8) pour déterminer les transformées de Fourier inverses. À la présente section, nous examinons deux exemples, les noyaux quartique et TH. Leur avantage est qu'il est facile de calculer leurs dérivées première et seconde, et que l'on peut obtenir une forme explicite de leurs transformées de Fourier inverses.

Exemple 1 : Noyau quartique

Nous commençons par considérer un noyau polynomial de degré quatre, c'est-à-dire quartique. L'imposition de toutes les contraintes (i) à (iv) donne la forme suivante :

$$A(\lambda) = \frac{15}{8\pi^4} (\lambda^4 - 2\pi^2 \lambda^2 + \pi^4),$$

$$\dot{A}(\lambda) = \frac{15}{8\pi^4} (4\pi^3 - 4\pi^2 \lambda), \text{ et}$$

$$\ddot{A}(\lambda) = \frac{15}{8\pi^4} (12\pi^2 - 4\pi^2 \lambda).$$

En prenant la transformée de Fourier inverse des noyaux de la pente et de la convexité (et leurs carrés), nous obtenons

$$\gamma_f(h) = \frac{1}{l} \int_{-\pi}^{\pi} f(\lambda) e^{ih\lambda} d\lambda, \quad (3)$$

une relation que nous utiliserons à plusieurs reprises dans la suite. Naturellement, cette relation entre γ_g^s et \mathcal{E} est vérifiée pour toute fonction intégrable g et non pas uniquement pour une densité spectrale. En outre, si nous désignons la matrice de Toeplitz associée à γ_g^s par $\Sigma(g)$, il s'ensuit que

$$\Sigma_k(g) = \gamma_g^s(f - k) = \frac{1}{l} \int_{-\pi}^{\pi} g(\lambda) e^{i(f-k)\lambda} d\lambda.$$

Or, d'après (2), f est une fonction d fois continûment dérivable si $\sum_{n=-\infty}^{\infty} |h|^n |\gamma_f(h)| < \infty$. Dans la suite de l'article, nous supposons que f est deux fois continûment dérivable (nous représenterons en abrégé cet espace de fonctions par C^2).

2.1 Mesures de la pente et de la convexité

La géométrie locale d'une fonction dans C^2 peut être décrite au moyen de ses dérivées première et seconde ; une mesure agrégée de ces dérivées s'obtient par intégration sur une bande de fréquences. Alternativement, on peut intégrer contre une fonction \mathcal{A} qui possède un support compact sur cette bande, à condition que \mathcal{A} fournisse une approximation appropriée pour l'intégration sur la bande. Nous désignons cette intégrale au moyen du dispositif général d'une fonctionnelle $\theta_{\mathcal{A}}$, où

$$\theta_{\mathcal{A}}(f) = \frac{1}{l} \int_{-\pi}^{\pi} \mathcal{A}(\lambda) f(\lambda) d\lambda. \quad (4)$$

La fonction \mathcal{A} sera appelée le « noyau » de cette fonctionnelle. Les mesures agrégées de la pente et de la convexité sont donc définies par $\theta_{\mathcal{A}}(f)$ et $\theta_{\mathcal{A}'}(f)$, où chaque point désigne une seule dérivée. Ces fonctionnelles donnent une mesure sommaire de la pente et de la convexité de f sur la bande $[l - \beta/2, l + \beta/2] \subset [0, \pi]$, et les noyaux correspondants seront par conséquent désignés par $\mathcal{A}_{\beta, l}$. Nous considérons les noyaux ayant les propriétés suivantes : i) $\mathcal{A}_{\beta, l}$ est une fonction dans C^2 sur $[-\pi, \pi]$; ii) $\mathcal{A}_{\beta, l}$ est nul en dehors de la bande $[l - \beta/2, l + \beta/2]$; iii) $\mathcal{A}_{\beta, l}$ est symétrique autour de l sur cette bande ; iv) $\mathcal{A}_{\beta, l}(l \pm \beta/2) = 0$. La condition (iii) garantit que l'emplacement du pic dans f ne sera pas modifié par l'emploi du noyau $\mathcal{A}_{\beta, l}$. Il convient de souligner que nous n'imposons pas que l'intégrale totale de $\mathcal{A}_{\beta, l}$ soit égale à l'unité, parce que plus tard, nous emploierons une normalisation qui tiendra compte automatiquement de la masse totale du noyau. Maintenant, au moyen de (iv) et d'une intégration par parties dans (4), nous obtenons

et

$$\mathcal{A}_{\beta, l}(\lambda) = \frac{4\pi^2}{2\pi} \mathcal{A}\left(\frac{\beta}{2\pi}(\lambda - l)\right),$$

$$\mathcal{A}_{\beta, l}(\lambda) = \frac{8\pi^3}{2\pi} \mathcal{A}\left(\frac{\beta}{2\pi}(\lambda - l)\right).$$

et

Plus tard, nous considérerons les carrés de ces noyaux, et leurs transformées de Fourier inverses correspondantes. Donc, en supposant que $[l - \beta/2, l + \beta/2] \subset [0, \pi]$, les carrés sont donnés par

$$\gamma_{\mathcal{A}_{\beta, l}}(h) = \exp\{ihl\} \gamma_{\mathcal{A}}(h\beta/2\pi), \quad (6)$$

où $\gamma_{\mathcal{A}_{\beta, l}}$ et $\mathcal{A}_{\beta, l}$ sont donnés par

$$\mathcal{A}_{\beta, l}(\lambda) = \frac{\beta}{2\pi} \mathcal{A}\left(\frac{\beta}{2\pi}(\lambda - l)\right),$$

et est nul en dehors de la bande de fréquence $[l - \beta/2, l + \beta/2]$. Manifestement, nous devons imposer que $\beta \leq 2\pi$ et $\beta \leq 2(\pi - l)$, pour que $[l - \beta/2, l + \beta/2] \subset [0, \pi]$; et le noyau $\mathcal{A}_{\beta, l}$ satisfait les conditions (i) à (iv). Notons que nous ne pouvons pas construire ces types de mesures si l est égal à 0 ou π . En modifiant les variables, nous voyons que

Ces formules sont commodes, parce qu'elles ne requièrent que la connaissance de f , et non celle de ses dérivées (en supposant que nous puissions calculer $\mathcal{A}_{\beta, l}$ et $\mathcal{A}_{\beta, l}'$). En nous inspirant de la littérature abondante sur les noyaux dans la régression non paramétrique et l'estimation de la densité spectrale, nous pouvons commencer par utiliser un noyau pair \mathcal{A} défini sur la bande $[-\pi, \pi]$ qui satisfait (i) et $\mathcal{A}(\pm\pi) = 0$. Alors, $\mathcal{A}_{\beta, l}$ est défini par

$$\theta_{\mathcal{A}_{\beta, l}}(f) = -\theta_{\mathcal{A}_{\beta, l}}(f) \quad (5)$$

$$\theta_{\mathcal{A}_{\beta, l}}(f) = \theta_{\mathcal{A}_{\beta, l}}(f).$$

négligative dans un voisinage raisonnablement grand de λ_0 . Cette réflexion aboutit au diagnostic proposé dans le présent article, c'est-à-dire des mesures agrégées de la pente et de la convexité de la densité spectrale, normalisées comme il convient. Mathématiquement, elles prennent la forme d'estimations d'un périodogramme lissé par la méthode du noyau, mais sans que la largeur de bande dépende de la taille de l'échantillon.

À la section 2, nous développons les notions mathématiques de cette méthode et les illustrons au moyen de deux noyaux soigneusement choisis. À la section 3, nous montrons comment formuler les estimateurs statistiques et comment tester les hypothèses statistiques concernant les pics. À la section 4, nous mettons la méthodologie à l'essai ; des simulations fournissent une description, en échantillon fini, de la taille et de la puissance de notre test. Nous démontrons en outre l'utilité de nos méthodes au moyen d'une étude de cas à grande échelle portant sur 130 séries chronologiques provenant du U.S. Census Bureau et de l'Organisation de coopération et de développement économiques (OCDE). Nous appliquons certains concepts tirés de la littérature sur les tests multiples (Hochberg 1988) pour combiner en un seul diagnostic des tests basés sur les fréquences individuelles. À la section 5, nous présentons nos conclusions, et à l'annexe, tous les théorèmes et preuves.

2. Mesure de la géométrie locale du spectre

En premier lieu, nous discutons de la géométrie de la densité spectrale (ou spectre) de la série chronologique étudiée. Comme point de départ, nous considérons des mesures de la pente et de la convexité du spectre qui sont entièrement déterministes (voir l'approche de Newton et Pagano 1983) ; plus tard, à la section 3, nous envisagerons des mesures statistiques. À la section 2.1, nous présentons la section 2.2, nous discutons de la pertinence de ces mesures pour la détection des pics, tandis qu'à la section 2.3, nous présentons, à titre d'exemples explicites, deux noyaux simples.

Supposons qu'après des transformations appropriées et une différenciation au besoin, X_1, X_2, \dots, X_n soit un échantillon issu d'un processus stochastique stationnaire de moyenne nulle. Nous utilisons la notation $X = (X_1, X_2, \dots, X_n)'$. La densité spectrale $f(\lambda)$, qui est bien définie à condition que la fonction d'autocovariance $\gamma_f(h)$ soit absolument sommable, est donnée par

$$(2) \quad f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_f(h) e^{-i\lambda h}$$

avec $i = \sqrt{-1}$ et $\lambda \in [-\pi, \pi]$. Il s'ensuit que la transformée de Fourier inverse donne

La présence de pics spectraux à des fréquences saisonnières dans une série désaisonnalisée peut signifier que les filtres saisonniers sont inadéquats – voir Soukup et Findley (1999) pour une discussion. Au minimum, les filtres de désaisonnalisation devraient éliminer la saisonnalité *non stationnaire* et tout effet périodique fixe, c'est-à-dire la saisonnalité qui, dans la série observée, contribue à la formation d'un pôle saisonnier dans le spectre. Cependant, la grande majorité des désaisonnalisateurs estiment qu'il est également souhaitable d'éliminer certains aspects de la saisonnalité *stationnaire* – d'où l'explosion des travaux de développement de filtres de désaisonnalisation fondés sur un

modèle (Bell et Hillmer 1984). La publication antérieure la plus importante à ce sujet est celle de Soukup et Findley (1999), qui proposent d'utiliser un spectre autorégressif pour trouver des pics « visuellement significatifs » – essentiellement, la valeur du spectre à chaque fréquence saisonnière (ou fréquence de jour ouvrable) est comparée aux valeurs voisines les plus proches, et est considérée comme un pic si l'écart est suffisamment grand. Cette méthode est appliquée à l'heure actuelle dans le programme X-12-ARIMA du U.S. Census Bureau (2002). L'une de ses limites tient au fait qu'elle ne comporte vraiment aucune composante statistique : la signification est non statistique – c'est-à-dire qu'elle n'est pas associée à un test d'hypothèse – et les seuls pour déterminer la « signification visuelle » sont établis de manière *ad hoc*. Dans le présent article, nous proposons un test de signification statistique pour la détection des pics qui peut donc être utilisé pour offrir des preuves statistiques supplémentaires de la présence d'un pic.

Un autre article apparenté est celui de Newton et Pagano (1983), qui élaborent des estimateurs convergents pour les maximums locaux du spectre. Notre approche est légèrement différente, en ce sens que nous connaissons déjà les fréquences d'intérêt (les six fréquences saisonnières), mais que nous cherchons à tester la présence d'un pic statistiquement significatif. Si nous considérons la vraie densité spectrale f comme une fonction lisse (qui peut être quantifiée au moyen d'une décroissance suffisamment rapide de la fonction d'autocovariance), un pic est une fréquence λ_0 telle que

$$(1) \quad f(\lambda_0) = 0 \quad \text{et} \quad f(\lambda_0) > 0,$$

où f et f' désignent les dérivées première et seconde. Manifestement, la dérivée seconde doit être négative avec une certaine signification pour que le concept ait un sens. Tout bien réfléchi, il semble naïf d'examiner la géométrie infinitésimale de f au point unique λ_0 , puisque n'importe quel petit pic dans une fonction monotone peut satisfaire (1), tout en étant dissocié des notions plus intuitives de ce qui constitue un pic. Par conséquent, il faut que la convexité soit

Un test non paramétrique pour la saisonnalité résiduelle

Tucker McElroy et Scott Holan¹

Résumé

La présence de pics dans le spectre d'un processus stationnaire signale l'existence de phénomènes périodiques stochastiques, tels que l'effet saisonnier. Nous proposons une mesure de ces pics spectraux et un test de détection de leur présence qui s'appuient sur l'évaluation de leur pente et de leur convexité agrégées. Notre méthode est élaborée de manière non paramétrique et peut donc être utile durant l'analyse préliminaire d'une série. Elle peut aussi servir à détecter la présence d'une saisonnalité résiduelle dans les données désaisonnalisées. Nous étudions le test diagnostique au moyen d'une simulation et d'une étude de cas à grande échelle portant sur des données provenant du U.S. Census Bureau et de l'Organisation de coopération et de développement économiques (OCDE).

Mots clés : Tests multiples ; estimation non paramétrique de la densité ; désaisonnalisation ; densité spectrale.

1. Introduction

La présence d'un pic dans le spectre d'un processus stationnaire est un indice de comportement périodique, tel que la saisonnalité ou un effet de jour ouvrable. Dans les domaines du génie et de l'économétrie, de nombreux auteurs s'intéressent à la détection de ces pics, car un pic spectral prononcé exerce une forte influence sur la dynamique du processus stochastique. Un pic est indicatif d'une gamme de fréquences dont la contribution à la variance globale du processus stochastique est relativement importante. Si la force du pic, évaluée par sa hauteur et sa largeur relativement aux valeurs voisines, est suffisamment prononcée, tout modèle de la dynamique du processus ne tenant pas compte des périodicités correspondantes sera spécifié incorrectement. Les ingénieurs ainsi que les économétristes voudraient vouloir extraire ou prévoir le signal, tâche qui, dans l'un et l'autre cas, est sensible à la présence de pics spectraux.

Par pic spectral, nous entendons une région de la densité spectrale dont la masse spectrale est supérieure à celle de ses voisines directes ; une définition plus précise est présentée plus loin. Dans le contexte des applications qui nous intéressent, les pics ont une hauteur finie et correspondent donc à des effets périodiques stochastiques dans un processus stationnaire. Par conséquent, notre principale préoccupation n'est pas la détection des effets périodiques fixes (déterministes) ni celle des phénomènes périodiques non stationnaires (quoique nous fassions certaines extensions à ce cas à la section 3.4) qui, dans les deux cas, correspondent à des pics spectraux de hauteur infinie. L'abondante littérature traitant de la détection des effets fixes est discutée dans Pritestley (1981) ; dans nos

applications, au lieu d'être fixes, les aspects périodiques évoluent au cours du temps.

Le présent article est axé sur l'application à la désaisonnalisation. Plus précisément, nous nous intéressons aux pics dits saisonniers, qui peuvent survenir aux fréquences saisonnières $\pi/6, 2\pi/6, 3\pi/6, 4\pi/6, 5\pi/6$ et $6\pi/6$ (en supposant que l'intervalle d'échantillonnage est mensuel). Dans le cas des statistiques fédérales, la détection de la saisonnalité et de la saisonnalité résiduelle pose un problème pratique important et le spectre est un outil logique pour l'effectuer. L'approche basée sur le domaine fréquentiel pour détecter et analyser la saisonnalité est très répandue, parce qu'elle offre un moyen tout naturel de visualiser un comportement quasi périodique. En fait, la saisonnalité est, informellement parlant, caractérisée par la présence d'au moins un pic saisonnier dans le spectre (Nerlove 1964). Les méthodes axées sur le domaine fréquentiel sont maintenant employées dans X-12-ARIMA (Findley, Monsell, Bell, Otto et Chen 1998) et font partie de TRAMO-SEATS (Maravall et Caporale 2004), qui sont les deux programmes de désaisonnalisation les plus répandus mis à la disposition du public. Notons que les méthodes du domaine fréquentiel peuvent être appliquées selon une approche paramétrique (c'est-à-dire fondée sur un modèle) ou non paramétrique. Nous mettons au point un diagnostic non paramétrique appelé pour déterminer l'efficacité de toute procédure de désaisonnalisation, qu'elle soit fondée sur un modèle ou non paramétrique. Comme l'ont souligné Findley, Monsell, Bell, Otto et Chen (1998), l'utilisation de fonctions périodiques fixes seulement pour modéliser la saisonnalité est habituellement inadéquate pour les données économiques (voir aussi la discussion dans Bell et Hillmer 1984).

1. Tucker McElroy, Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100. Courriel : tucker_s.mcelroy@census.gov; Scott Holan, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100. Courriel : holans@missouri.edu.

$$0 = E(T_r(b)) \quad \text{ou } K_r = ((K_{rst})).$$

D'où,

$$E(b-b) = \frac{1}{2} \Sigma^{-1} \begin{pmatrix} tr(\Sigma^{-1} K_1) \\ \vdots \\ tr(\Sigma^{-1} K_p) \end{pmatrix} + O(n_r^{-1/2}).$$

Puisque $n_r = O_p(k)$, le théorème s'ensuit.

Bibliographie

Botman, S.L., Moore, T.F., Mortality, C.L. et Parsons, V.L. (2000). Design and estimation for the National Health Interview Survey, 1995-2004. *Vital and Health Statistics*, 2, 130.

Cox, D.R., et Snell, E.J. (1968). A general distribution of residuals (avec discussion). *Journal of the Royal Statistical Society, Séries B*, 30, 248-275.

Ghosh, M., et Maity, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, 91, 95-112.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Gelman, A., Carlin, J., Stern, H. et Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A., et Rubin, D. (1992). Inference from iterative simulation using multiple sequences (avec discussion). *Statistical Science*, 457-511.

Godambe, V.P., et Thompson, M.E. (1989). An extension of quasi-likelihood estimation (avec discussion). *Journal of Statistical Planning and Inference*, 22, 137-152.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10, 65-80.

Morris, C. (1983). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 11, 515-529.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Sarkar, S., et Ghosh, M. (1998). Empirical Bayes estimation of local area means for NEF-QVF superpopulations. *Sankhya, Séries B*, 60, 464-487.

Tanner, M.A. (1996). *Tools for Statistical Inference*. New York : Springer.

En outre, posons que

$$K_{rst} = E \left[\frac{\partial^2 T_r(b)}{\partial b_s \partial b_t} \right] = \frac{\partial}{\partial b_t} \sum_{j=1}^l \sum_{i=1}^n (1 - m_{ij}(b))(1 - m_{ij}(b)) x_{ijr} x_{ijst} = - \sum_{j=1}^l \sum_{i=1}^n (1 - 2m_{ij}(b)) m_{ij}(b) (1 - m_{ij}(b)) x_{ijr} x_{ijst}. \quad (A.12)$$

Donc, nous avons

$$\sum_k \sigma_{rs} E(\hat{b}_s - b_s) = \sum_p \sum_{s=1}^s \sigma_{rs} K_{rst}, \quad r = 1, \dots, p.$$

En notation matricielle, nous obtenons

$$\Sigma E(b-b) = \frac{1}{2} \begin{pmatrix} tr(\Sigma^{-1} K_1) \\ \vdots \\ tr(\Sigma^{-1} K_p) \end{pmatrix}$$

Donc, $b - b = \Sigma^{-1}(T(b) + O^p(n_T^{-1})) = m_{ij}^{(b)}$. Puisque $A^{(Y)} = m_{ij}^{(b)}$ (1 - $m_{ij}^{(b)}$), $A^{(T)}(b) = \Sigma(b)$. D'où $E[(b - b)^T] = \Sigma^{-1}(b) + O(n_T^{-3/2})$. En outre, dans (8.5), nous avons $E(x_{ij}^T(b - b)^2) = tr(x_{ij}^T \Sigma^{-1}(b)) + O^p(n_T^{-1})$. Conséquemment, par (A.4) et (A.5), nous obtenons l'approximation

$$E[m_{ij}^{(b)}(b) - m_{ij}^{(b)}(b)]^2 = m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))^2 x_{ij}^T \Sigma^{-1}(b) x_{ij} \quad (A.7)$$

qui est correcte jusqu'à l'ordre $O(n_T^{-1})$ en vertu de notre hypothèse.

Souignons que le terme négligé

$$E[x_{ij}^T(b - b)(1 - 2m_{ij}^{(b)}(b))(b - b)^T x_{ij} x_{ij}^T(b - b)] = O(n_T^{-3/2})$$

puisque

$$E[x_{ij}^T(b - b)(1 - 2m_{ij}^{(b)}(b))(b - b)^T x_{ij} x_{ij}^T(b - b)] = (1 - 2m_{ij}^{(b)}(b))E[x_{ij}^T(b - b)^T x_{ij}]$$

$$= (1 - 2m_{ij}^{(b)}(b))E[x_{ij}^T(b - b)]^2$$

$$= O(n_T^{3/2}).$$

De même, notons que $E[x_{ij}^T(b - b)]^4 = O(n_T^{-2})$ et

$$Em_{ij}^{(b)}(b) - m_{ij}^{(b)}(b) = 0$$

$$= m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))(1 - m_{ij}^{(b)}(b))x_{ij}^T \Sigma^{-1}(b)x_{ij}$$

$$+ O(n_T^{-3/2}). \quad (A.8)$$

Cela nous mène à

$$E\left[\sum_{j=1}^f w_{ij}^{(b)} (d_{ij}^{EB} - d_{ij}^{(b)})^2\right] = \frac{\lambda^2}{2} \sum_{j=1}^f w_{ij}^{(b)} m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))x_{ij}^T \Sigma^{-1}(b)x_{ij}$$

$$+ \sum_{1 \leq f \neq j \leq n_{ij}} w_{ij}^{(b)} m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))m_{ij}^{(b)}(b) \left[\frac{\lambda^2}{2} (1 + \gamma) \right]$$

$$= \frac{\lambda^2}{2} (1 + \gamma) \sum_{j=1}^f w_{ij}^{(b)} m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))x_{ij}^T \Sigma^{-1}(b)x_{ij} + O(n_T^{-3/2})$$

$$\times \Sigma^{-1}(b) \sum_{j=1}^f w_{ij}^{(b)} m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))x_{ij} + O(n_T^{-3/2}). \quad (A.9)$$

Puisque $\Sigma^{-1}(b) = O(k^{-1})$, et $n_T = O(k)$ en vertu de notre hypothèse, le théorème découle de (A.2) et (A.9).

Preuve du théorème 2. Nous commençons par noter que $b = b + O^p(n_T^{-1/2}) = b + O^p(k^{-1})$ et $\Sigma^{-1}(b) = O(k^{-1})$. D'où, le deuxième terme du deuxième membre de (8.7) est approximé par

$$c \left[\sum_{j=1}^f w_{ij}^{(b)} m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))x_{ij}^T \right]$$

($c = \lambda^2/(1 + \gamma)^2$) qui est exact jusqu'à l'ordre $O(k^{-1})$. Cependant, si nous estimons $m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))$ simplement par $m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))$, nous ignorons le terme d'ordre $O(k^{-1})$. Donc, nous avons besoin d'une approximation prudente du biais $E(b - b)$ pour obtenir l'approximation souhaitée. Pour cela, nous nous inspirons de Cox et Snell (1968).

$$E[m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b))] =$$

$$= E[(m_{ij}^{(b)}(b) + m_{ij}^{(b)}(b))(1 - m_{ij}^{(b)}(b)) + m_{ij}^{(b)}(b) - m_{ij}^{(b)}(b)]$$

$$= m_{ij}^{(b)}(b)(1 - m_{ij}^{(b)}(b)) + (1 - 2m_{ij}^{(b)}(b))E[m_{ij}^{(b)}(b)]$$

$$- E[m_{ij}^{(b)}(b) - m_{ij}^{(b)}(b)].$$

Maintenant, de nouveau par développement en série de Taylor en deux étapes, nous avons

$$E[m_{ij}^{(b)}(b) - m_{ij}^{(b)}(b)] = \left[\frac{\partial}{\partial b} m_{ij}^{(b)}(b) \right]^T E(b - b)$$

$$+ \frac{1}{2} E \left[(b - b)^T \frac{\partial^2 m_{ij}^{(b)}(b)}{\partial b \partial b^T} (b - b) \right] + O(n_T^{-3/2}).$$

Afin de trouver $E(b - b)$, nous procédons comme il suit. Nous commençons par le développement en série de Taylor de deuxième ordre

$$0 = T_r(b) = T_r(b) + \sum_{s=1}^s (\hat{b}_s - b_s) \frac{\partial}{\partial b_s} T_r(b)$$

$$+ \frac{1}{2} \sum_{s=1}^s \sum_{p=1}^s (\hat{b}_s - b_s)(\hat{b}_p - b_p) \frac{\partial^2 T_r(b)}{\partial b_s \partial b_p} + O(n_T^{-3/2}).$$

En prenant les espérances et en suivant Cox et Snell (1968),

$$E(p_B^{ij} - p^{ij})^z = E \left[\frac{1}{1} x^{ij} + \frac{\lambda + 1}{\lambda} m^{ij}(b) - p^{ij} \right]^z$$

$$= E \left[\frac{\lambda + 1}{1} (y^{ij} - p^{ij}) + \frac{\lambda}{\lambda} (m^{ij}(b) - p^{ij}) \right]^z$$

$$= \frac{1}{1} E(y^{ij} - p^{ij})^z + \frac{\lambda}{\lambda} E(m^{ij}(b) - p^{ij})^z$$

$$+ \frac{\lambda}{2\lambda} E(y^{ij} - p^{ij})(m^{ij}(b) - p^{ij})$$

$$= \frac{1}{1} E(p^{ij} (1 - p^{ij})) + \frac{\lambda}{\lambda} A(p^{ij}) + 0$$

$$= \frac{1}{1} \left(\frac{\lambda + 1}{\lambda} m^{ij}(b) (1 - m^{ij}(b)) \right)$$

$$+ \frac{\lambda}{\lambda} \left(\frac{\lambda + 1}{m^{ij}(b) (1 - m^{ij}(b))} \right)$$

$$= \frac{\lambda m^{ij}(b) (1 - m^{ij}(b))}{(\lambda + 1)^2},$$

de sorte que

$$\left[\sum_{j=1}^n w_j (p_B^{ij} - p^{ij}) \right]^z$$

$$= \frac{\lambda}{\lambda} \sum_{j=1}^n w_j^2 m^{ij}(b) (1 - m^{ij}(b)).$$

(A.2)

Enfin, nous calculons,

$$\left[\sum_{j=1}^n w_j (p_B^{ij} - p^{ij}) \right]^z$$

$$= \frac{\lambda}{\lambda} \left[\sum_{j=1}^n w_j^2 m^{ij}(b) (1 - m^{ij}(b)) \right]^z$$

$$= \frac{\lambda}{\lambda} \left[\sum_{j=1}^n w_j^2 m^{ij}(b) (1 - m^{ij}(b)) \right]^z$$

$$+ \sum_{j \leq n} \sum_{j \neq n} w_j w_{j'} (m^{ij}(b) - m^{ij'}(b))$$

$$\left[\sum_{j=1}^n w_j (m^{ij}(b) - m^{ij'}(b)) \right]^z.$$

(A.3)

Par développement en série de Taylor en deux étapes,

$$= -\Sigma(b).$$

(A.6)

$$= -X^T M(I - M)X$$

$$= - \sum_{j=1}^n \sum_{k=1}^n m_{jk}^{ij}(b) (1 - m_{jk}^{ij}(b)) x_{jk}^{ij} x_{jk}^{ij}$$

$$\nabla T(b) = - \sum_{j=1}^n \sum_{k=1}^n \left(\frac{\partial}{\partial m_{jk}^{ij}(b)} \right) x_{jk}^{ij}$$

Afin de trouver $E[(b - b)^T]$, nous procédons comme il suit. Soit $T(b) = \sum_{j=1}^n \sum_{k=1}^n (y_{jk}^{ij} - m_{jk}^{ij}(b)) x_{jk}^{ij}$ de sorte que $T(b) = 0$. Par développement en série de Taylor en une étape, où $0 = T(b) = T(b) + [\nabla T(b)]^T (b - b) + O_p(n^{-1})$, ou

$$= E[x_{jk}^{ij} x_{jk}^{ij} E(b - b)^T].$$

(A.5)

$$E[x_{jk}^{ij} x_{jk}^{ij}]^2 = E[(b - b)^T x_{jk}^{ij} x_{jk}^{ij} (b - b)]$$

observons que

Le premier terme négligé est $O_p(\|b - b\|)$. De Sarkar et Ghosh (1998), il découle que $b - b$ est asymptotiquement de loi $N(0, \Sigma^{-1}(b))$, où $\Sigma(b) = O_p(k)$, il découle que Sous l'hypothèse que $\Sigma(b) = O_p(k)$, il découle que $\Sigma^{-1}(b) = O_p(k^{-1})$. Donc, $\|b - b\| = O_p(k^{-1/2})$. D'où, le premier terme négligé est $O_p(k^{-3/2})$. Ensuite, nous

$$(b - b)^T x_{jk}^{ij} x_{jk}^{ij} (b - b) \left[\right]^2.$$

(A.4)

$$= m_{jk}^{ij}(b) (1 - m_{jk}^{ij}(b))^2 E \left[x_{jk}^{ij} x_{jk}^{ij} \right]^2 + \frac{1}{2} (1 - 2m_{jk}^{ij}(b))$$

$$m_{jk}^{ij}(b) (1 - m_{jk}^{ij}(b)) (1 - 2m_{jk}^{ij}(b)) x_{jk}^{ij} x_{jk}^{ij} (b - b) \left[\right]^2$$

$$= E \left[m_{jk}^{ij}(b) (1 - m_{jk}^{ij}(b)) x_{jk}^{ij} x_{jk}^{ij} (b - b) + \frac{1}{2} (b - b)^T \right]$$

$$E[m_{jk}^{ij}(b) (1 - m_{jk}^{ij}(b))]^2$$

Notant que $(\partial^2 m_{jk}^{ij}(b)) / (\partial b \partial b^T) = (1 - 2m_{jk}^{ij}(b)) m_{jk}^{ij}(b) (1 - m_{jk}^{ij}(b)) x_{jk}^{ij} x_{jk}^{ij}$, il s'ensuit que

$$(b - b) + \frac{1}{2} (b - b)^T \frac{\partial^2 m_{jk}^{ij}(b)}{\partial b \partial b^T} (b - b).$$

$$m_{jk}^{ij}(b) = m_{jk}^{ij}(b) + \left(\frac{\partial}{\partial m_{jk}^{ij}(b)} \right)^T$$

7. Résumé, futurs travaux et discussion

L'estimation de la proportion de personnes non assurées, particulièrement parmi les groupes minoritaires, est définitivement un problème très important, susceptible d'avoir une incidence sur l'élaboration des politiques des organismes fédéraux et des États. Nous venons de commencer à aborder cette question très importante et fournissons des estimations pour petits domaines bayésiennes empiriques et hiérarchiques pour la sous-population asiatique, subdivisée par croisement de l'âge, du sexe et d'autres caractéristiques démographiques. Nous discutons également de l'adéquation de l'ajustement de notre modèle au moyen de la valeur p prédictive à posteriori. Néanmoins, beaucoup de travaux restent à faire. En particulier, nous voulons étendre les présents résultats à l'analyse de données binaires bivariées et multivariées.

Comme l'a souligné un examinateur, la présente analyse ne tient pas compte de la mise en grappes des ménages dans le calcul de la vraisemblance, puisque l'enquête originale est une enquête-ménage et que, très catégoriquement, la couverture par une assurance est corrélée à l'intérieur des ménages. Cependant, notre modèle hypothétique n est que conditionnellement indépendamment, sachant les covariables les effets aléatoires. Une fois que nous avons attribué des distributions aux coefficients de régression et aux composantes de variance, la dépendance est intégrée automatiquement dans le modèle final, au niveau de l'unité ainsi que du haut, nous avons testé l'adéquation du modèle hiérarchique bayésien au moyen des valeurs p prédictives à posteriori.

Enfin, il convient de souligner que ces travaux de recherche ne sont présentés ici qu'à titre illustratif. L'application de cette méthode dans un contexte d'élaboration de politiques nécessiterait un examen plus approfondi des méthodes et du respect des normes institutionnelles relatives à la diffusion de politiques officielles.

Remerciements

Nous remercions le rédacteur adjoint et deux examinateurs de leurs commentaires constructifs au sujet d'ébauches antérieures de l'article. Les travaux de recherche du premier auteur ont été financés en partie par les bourses SES-0317589 et SES-0631426 de la NSF. Ceux du quatrième auteur ont été financés en partie par la bourse SES-0318184 de la NSF. Les travaux ont également été financés partiellement par les NCHS/CDC sous le projet numéro 282286285 intitulé « Model-Based Subdomain Estimates ». Les constatations et les conclusions du présent article sont celles des auteurs et ne reflètent pas nécessairement les

$$E\mathbb{Q}M(\underline{\mu}_{EB}^{tw})$$

Preuve du théorème 1.

Annexe

opinions du National Center for Health Statistics des Centers for Disease Control and Prevention.

$$= E(\underline{\mu}_{EB}^{tw} - \underline{\mu}_{EB}^{tw})^2 = E\left(\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij})\right)^2$$

$$= E\left(\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij} - p_{ij} + p_{ij})\right)^2$$

$$= E\left(\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij})\right)^2 + E\left(\sum_{j=1}^n w_{ij} (p_{ij} - p_{ij})\right)^2$$

$$+ 2E\left(\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij})\right)\left(\sum_{j=1}^n w_{ij} (p_{ij} - p_{ij})\right).$$

En notant que $E(p_{ij} | y) = p_{ij}^*$,

$$\left[\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij})\right]\left(\sum_{j=1}^n w_{ij} (p_{ij} - p_{ij})\right) = 0.$$

$$= E\left[\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij})\right] \times E\left(\sum_{j=1}^n w_{ij} (p_{ij} - p_{ij}) | y\right) = 0.$$

D'où,

$$E\mathbb{Q}M(\underline{\mu}_{EB}^{tw}) = E\left(\sum_{j=1}^n w_{ij} (p_{EB}^{ij} - p_{ij})\right)^2$$

$$(A.1) \quad + E\left(\sum_{j=1}^n w_{ij} (p_{ij} - p_{ij})\right)^2.$$

Mais

$$E\left(\sum_{j=1}^n w_{ij} (p_{ij} - p_{ij})\right)^2 = \sum_{j=1}^n w_{ij}^2 E(p_{ij} - p_{ij})^2.$$

Ensuite, nous calculons

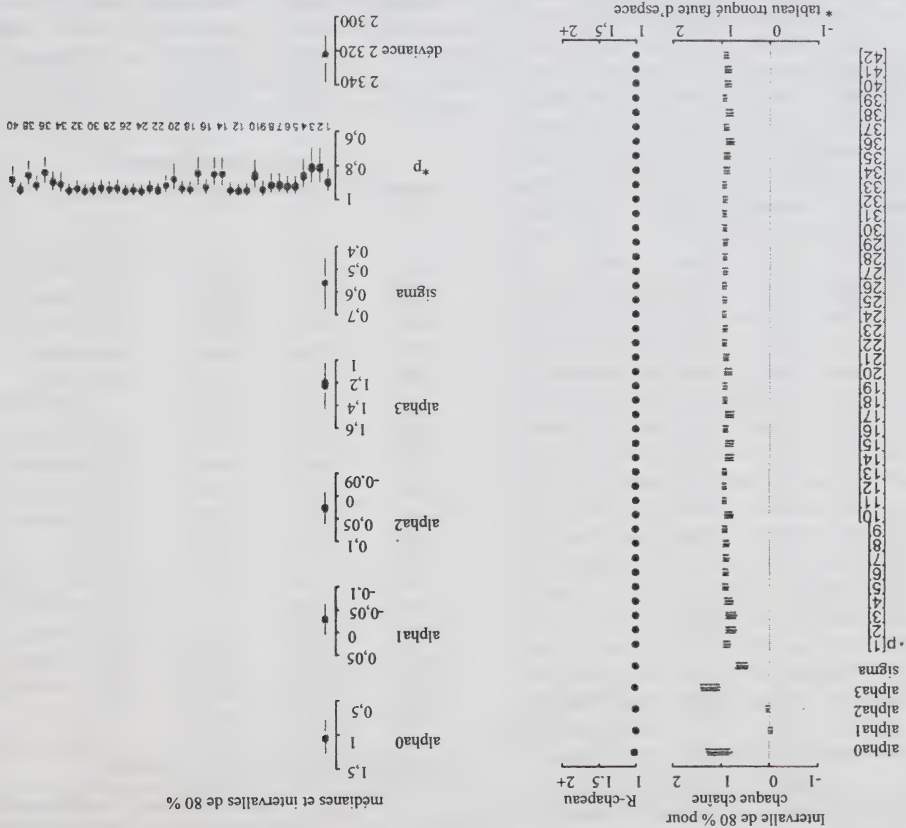


Figure 1 Model Bugs à « asian_model.bug », 5 chaînes, chacune avec 1 000 itérations

où $D(y, \theta)$ est la déviance calculée aux valeurs estimées des paramètres et $D(y, \theta^{(i)})$ est la déviance estimée en utilisant la simulation a posteriori. Pour des détails, voir Gelman et coll. (2004). Pour notre analyse HB, $p_d = 56,75$ et $DIC = 2414,41$. Habituellement, les valeurs p_d et DIC sont utilisées comme critères d'ajustement du modèle et pour choisir le modèle ayant le meilleur pouvoir prédictif. Donc, nous ajustons aussi le modèle de régression logistique simple (modèle courant sans aucun effet aléatoire), ce qui signifie qu'il n'y a pas de groupement de données, et les valeurs estimées de p_d et DIC sont 22,60 et 2379,55, respectivement. La valeur p correspondante est 0,3848. Donc le modèle proposé semble être raisonnablement bien ajusté aux données.

Pour le modèle de régression logistique hiérarchique bayésien, la valeur p estimée est 0,4216 pour $(c, d) = (0,02, 0,02)$. Les autres choix de (c, d) produisent des valeurs semblables. Une valeur p supérieure à 0,3 est habituellement traitée comme un bon ajustement. Donc, la méthode HJB proposée semble donner de bons résultats dans la situation décrite ici.

Nous avons également calculé la valeur $p_d = \text{var}(\text{deviance})/2$ et le critère d'information de déviance (DIC) pour $\text{deviance information criterion}$ ou la déviance prédictive estimée. La valeur p_d peut être vue comme le nombre de paramètres non contraints dans le modèle, où un paramètre prend la valeur 1 s'il fait partie du modèle original (distribution des données) et la valeur 0 s'il est associé à toute loi a priori. Le DIC est estimé sous la forme

$$DIC = 2D(y, \theta^{(i)}) - D(y, \theta)$$

Techniques d'enquête, juin 2009

Tableau 2

Proportions n'ayant pas d'assurance-maladie, selon le groupe d'âge et le groupe d'Asiatiques en 2000

0 à 17 ans					
Directes		HB	EB ($\lambda = 0,5$) EB ($\lambda = 1$)		
Total	0,120	0,126	0,131	0,114	0,137
Asiatique 1	0,087	0,097	0,105	0,083	0,114
Asiatique 2	0,046	0,062	0,071	0,059	0,083
Asiatique 3	0,113	0,117	0,114	0,114	0,114
Asiatique 4	0,165	0,165	0,171	0,171	0,175
18 à 64 ans					
Total	0,177	0,172	0,168	0,164	0,164
Asiatique 1	0,162	0,160	0,160	0,159	0,159
Asiatique 2	0,137	0,137	0,135	0,134	0,134
Asiatique 3	0,150	0,147	0,141	0,137	0,137
Asiatique 4	0,219	0,208	0,203	0,195	0,195
65 ans et plus					
Total	0,063	0,080	0,103	0,123	0,123
Asiam 1	0,083	0,097	0,123	0,143	0,143
Asiam 2	0,021	0,043	0,064	0,085	0,085
Asiam 3	0,119	0,126	0,136	0,145	0,145
Asiam 4	0,055	0,075	0,100	0,123	0,123

6. Diagnostics et mise en œuvre du modèle hiérarchique bayésien

Pour la mise en œuvre et les diagnostics de convergence de l'échantillonneur de Gibbs, nous nous sommes inspirés de Gelman et Rubin (1992). En particulier, nous nous sommes servis de cinq chaînes, chacune de taille 1 000, avec une période d'apprentissage initiale de 1 000 itérations. Nous avons vérifié la convergence des facteurs de réduction d'échelle possibles et elle semblait être très proche de 1 unité (= 1 à la convergence) pour chacun des paramètres. Plusieurs autres critères diagnostiques décrits dans la littérature ont été appliqués au moyen du logiciel CODA. Une sortie partielle est présentée à la figure 1. Le côté gauche montre le chevauchement des cinq chaînes parallèles et le côté droit, l'inférence a posteriori pour chaque paramètre et la déviance (-2 log vraisemblance). Pour une description détaillée du logiciel que nous avons utilisé, nous renvoyons le lecteur à l'annexe C de Gelman, Carlin, Stern et Rubin (2004).

Un moyen bayésien de vérifier la qualité de l'ajustement d'un modèle aux données consiste à tirer des valeurs simulées de la loi prédictive a posteriori des données observées et de comparer ces échantillons aux données répliquées. Un écart important entre les données générées et les données observées indique un manque d'ajustement du modèle. À l'instar de Gelman et coll. (2004), nous avons calculé les valeurs p bayésiennes pour vérifier l'adéquation des modèles bayésiens proposés. La justification générale de

ce genre de calcul est la suivante. Soit y le vecteur de données observées, ξ le vecteur de paramètres inconnus, $f(y|\xi)$ la densité de y sachant ξ et $\Pi(\xi|y)$, la densité a posteriori de ξ sachant y . Supposons que nous avons tiré les échantillons $\xi^{(1)}, \dots, \xi^{(R)}$ de cette loi a posteriori en utilisant une simulation MCMC. Simulons maintenant R répliques hypothétiques des données, disons $y^{(1)}, \dots, y^{(R)}$, sachant le $\xi^{(l)}$ simulé. Si le modèle est raisonnablement précis, ces répliques hypothétiques devraient être similaires aux données observées y . Formellement, nous faisons cela en commençant par choisir une variable de divergence, disons $d(y, \xi)$, qui a une valeur extrême si les données y sont entièrement en désaccord avec le modèle donné. Puis, nous estimons une valeur p par la proportion de cas dans lesquels la variable de divergence simulée excède la valeur réalisée de cette variable. Donc, la valeur p estimée (habituellement appelée valeur p prédictive a posteriori) est égale à $R^{-1} \sum_{l=1}^R I[d(y^{(l)}, \xi^{(l)}) \geq d(y, \xi^{(l)})]$ où I est la fonction indicatrice habituelle. Pour vérifier la qualité de l'ajustement du modèle, nous pouvons utiliser un diagramme de dispersion des valeurs réalisées $d(y, \xi^{(l)})$ sur la même échelle. Si la moitié environ des points du diagramme de dispersion se trouvent au-dessus de la droite à 45 degrés et l'autre moitié, sous celle-ci, l'ajustement est bon. Autrement dit, pour les grands échantillons, la valeur p estimée s'écartera peu d'une moitié. Naturellement, nous pouvons également exécuter une analyse graphique en utilisant différents diagrammes pour divers sous-groupes, ce qui permettrait de visualiser un échec local éventuel du modèle qui, autrement, pourrait être obscurci dans le diagramme de dispersion agrégé.

Plusieurs choix sont possibles pour la variable de divergence d . Dans le cas qui nous occupe, nous en avons considéré une en particulier. Notant que $E(X|p_j) = P_j = \exp(\theta_j)/(1 + \exp(\theta_j))$, nous pouvons considérer le carré des résidus standardisés $((y_j^{(l)} - P_j)/(P_j(1 - P_j)))$, où $P_j^{(l)} = \exp(\theta_j^{(l)})/(1 + \exp(\theta_j^{(l)}))$ sont les valeurs générées des p_j à partir de la l^{e} itération. Alors, la variable de divergence d est

$$d(y, p^{(l)}) = \sum_{j=1}^n \sum_{l=1}^{95} \frac{d_{(j)}^{(l)}}{d_{(j)}^{(l)} + 1}$$

$$d(y, p^{(l)}) = \sum_{j=1}^n \sum_{l=1}^{95} \frac{d_{(j)}^{(l)}}{d_{(j)}^{(l)} + 1}$$

Manifestement, il existe d'autres choix possibles de d . Gelman et Chosh (1998) ont proposé un certain nombre de mesures de divergence et étudié leurs propriétés.

Tableau 1 (suite)

Estimations pour petits domaines des proportions d'Asiatiques non assurés : année 2000

Domaine	n_i	Directe	Moyenne	1997-1999	HB	$\hat{\epsilon}$ -L (HB)	EB	$\lambda = 0.5$	EB	$\lambda = 1$	$\hat{\epsilon}$ -L (EB)	$\lambda = 0.5$	EB	$\lambda = 1$	$\hat{\epsilon}$ -L (EB)
54	18	0.202	0.169	0.195	0.031	0.188	0.019	0.020	0.019	0.019	0.020	0.019	0.019	0.019	0.020
55	74	0.094	0.115	0.102	0.015	0.102	0.019	0.020	0.019	0.019	0.020	0.019	0.019	0.019	0.020
60	8	0.112	0.116	0.120	0.044	0.132	0.082	0.028	0.059	0.059	0.030	0.063	0.063	0.044	0.049
63	33	0.055	0.093	0.069	0.020	0.073	0.032	0.032	0.051	0.051	0.030	0.063	0.063	0.044	0.049
64	28	0.105	0.275	0.112	0.024	0.115	0.120	0.032	0.032	0.032	0.034	0.063	0.063	0.044	0.049
65	33	0.126	0.133	0.129	0.021	0.126	0.126	0.032	0.032	0.032	0.034	0.063	0.063	0.044	0.049
70	2	0.361	0.000	0.331	0.098	0.299	0.268	0.119	0.077	0.077	0.082	0.165	0.165	0.044	0.049
72	2	0.000	0.182	0.045	0.101	0.157	0.236	0.155	0.051	0.051	0.055	0.165	0.165	0.044	0.049
73	45	0.271	0.144	0.256	0.026	0.256	0.249	0.028	0.051	0.051	0.055	0.165	0.165	0.044	0.049
74	10	0.000	0.044	0.024	0.034	0.034	0.051	0.051	0.051	0.051	0.055	0.165	0.165	0.044	0.049
75	83	0.149	0.097	0.150	0.016	0.160	0.166	0.020	0.021	0.021	0.021	0.055	0.055	0.044	0.049
76	59	0.113	0.205	0.120	0.018	0.128	0.136	0.023	0.023	0.023	0.024	0.055	0.055	0.044	0.049
77	68	0.338	0.224	0.313	0.025	0.302	0.284	0.023	0.023	0.023	0.024	0.055	0.055	0.044	0.049
78	39	0.098	0.138	0.103	0.020	0.102	0.104	0.026	0.026	0.026	0.028	0.055	0.055	0.044	0.049
79	122	0.110	0.163	0.117	0.013	0.125	0.133	0.016	0.016	0.016	0.017	0.055	0.055	0.044	0.049
80	125	0.308	0.314	0.281	0.020	0.262	0.239	0.016	0.016	0.016	0.017	0.055	0.055	0.044	0.049
81	7	0.000	0.000	0.029	0.043	0.066	0.099	0.065	0.065	0.065	0.069	0.051	0.051	0.044	0.049
82	12	0.000	0.045	0.025	0.032	0.047	0.070	0.048	0.048	0.048	0.051	0.051	0.051	0.044	0.049
83	13	0.049	0.017	0.068	0.035	0.088	0.108	0.050	0.050	0.050	0.053	0.053	0.053	0.044	0.049
84	4	0.000	0.061	0.028	0.056	0.060	0.091	0.088	0.088	0.088	0.093	0.093	0.093	0.044	0.049
85	32	0.189	0.113	0.193	0.027	0.217	0.231	0.035	0.035	0.035	0.037	0.037	0.037	0.044	0.049
86	10	0.136	0.056	0.137	0.036	0.127	0.123	0.051	0.051	0.051	0.054	0.054	0.054	0.044	0.049
87	52	0.192	0.098	0.185	0.021	0.184	0.180	0.024	0.024	0.024	0.026	0.026	0.026	0.044	0.049
88	65	0.153	0.210	0.155	0.018	0.162	0.166	0.022	0.022	0.022	0.024	0.024	0.024	0.044	0.049
89	71	0.285	0.210	0.265	0.022	0.256	0.256	0.022	0.022	0.022	0.023	0.023	0.023	0.044	0.049
90	57	0.086	0.146	0.095	0.017	0.102	0.110	0.022	0.022	0.022	0.024	0.024	0.024	0.044	0.049
91	153	0.149	0.167	0.150	0.011	0.156	0.160	0.014	0.014	0.014	0.015	0.015	0.015	0.044	0.049
92	138	0.308	0.285	0.283	0.020	0.266	0.244	0.015	0.015	0.015	0.017	0.017	0.017	0.044	0.049
93	10	0.000	0.000	0.030	0.041	0.073	0.110	0.059	0.059	0.059	0.063	0.063	0.063	0.044	0.049
94	16	0.067	0.015	0.081	0.029	0.090	0.101	0.0101	0.0101	0.0101	0.044	0.044	0.044	0.044	0.049
95	18	0.108	0.018	0.123	0.032	0.145	0.163	0.046	0.046	0.046	0.049	0.049	0.049	0.044	0.049
96	14	0.111	0.087	0.125	0.039	0.160	0.185	0.050	0.050	0.050	0.053	0.053	0.053	0.044	0.049

Les estimations HB de la proportion de personnes non assurées pour le groupe des Asiatiques varient de 2 % à 35 % pour les petits domaines, à l'exclusion du domaine 69. Il faut reconnaître que, pour ce dernier, les estimations EB et HB sont affectées de façon considérable par la petite taille d'échantillon. Nous présentons également les erreurs-types associées aux estimations HB, ainsi que la racine des erreurs quadratiques moyennes approximatives estimées des estimations EB. L'approche proposée répond largement à la critique valide selon laquelle les estimations EB naïves des erreurs-types (qui ne tiennent pas compte du terme $O(K^{-1})$) sont habituellement des sous-estimations. Nous donnons également une colonne contenant la moyenne de trois ans des estimations directes pour 1997 à 1999. Celle-ci a principalement pour but d'examiner si les domaines pour lesquels l'estimation directe était nulle en 2000 possèdent la même caractéristique pour d'autres années, ainsi que de comparer les estimations directes et HB à ces estimations plutôt qu'aux estimations directes. Il s'avère qu'à très peu d'exceptions près, la moyenne pour 1997 à 1999 concorde peu aux estimations directes. Cependant, l'estimation directe demeure nulle pour le domaine 69.

Le tableau 2 donne un sommaire des proportions de personnes non assurées dans les trois groupes d'âge, à savoir 0 à 17 ans, 18 à 64 ans et 65 ans et plus, individuellement pour les Chinois (Asiatique 1), les Philippins (Asiatique 2), les Indiens d'Asie (Asiatique 3) et les autres Asiatiques (Asiatique 4). Il s'avère qu'à ce niveau plus élevé d'aggrégation, les estimations EB et HB pour petits domaines sont assez proches des estimations directes, sauf peut-être pour le groupe des 65 ans et plus. Les résultats semblent assez satisfaisants, puisqu'à ce niveau d'aggrégation, les comparaisons.

Tableau 1
Estimations pour petits domaines des proportions d'Asiatiques non assurés : année 2000

Domaine	n_i	Directe	Moyenne 1997 à 1999	HB	É-L (HB)	EB	EB	É-L (EB)	É-L (EB)
1	10	0,126	0,034	0,133	0,043	0,148	0,158	0,057	0,060
2	0	-	0,085	-	-	-	-	-	-
3	24	0,063	0,016	0,074	0,025	0,076	0,082	0,037	0,039
4	28	0,146	0,105	0,150	0,027	0,163	0,171	0,041	0,043
5	20	0,138	0,265	0,143	0,032	0,153	0,160	0,043	0,046
6	17	0,112	0,124	0,120	0,032	0,134	0,144	0,019	0,021
7	78	0,097	0,100	0,104	0,015	0,107	0,112	0,022	0,024
8	66	0,274	0,229	0,253	0,023	0,240	0,224	0,072	0,076
9	5	0,000	0,000	0,164	0,061	0,160	0,154	0,078	0,082
10	6	0,000	0,000	0,033	0,051	0,082	0,123	0,070	0,074
11	7	0,000	0,084	0,032	0,047	0,090	0,134	0,054	0,057
12	11	0,335	0,000	0,302	0,056	0,275	0,245	0,060	0,064
13	7	0,134	0,061	0,134	0,045	0,130	0,128	0,103	0,110
14	2	0,000	0,151	0,020	0,064	0,026	0,039	0,031	0,033
15	27	0,000	0,104	0,023	0,023	0,035	0,052	0,032	0,034
16	29	0,113	0,191	0,119	0,024	0,123	0,127	0,033	0,035
17	27	0,120	0,223	0,127	0,025	0,141	0,152	0,044	0,047
18	14	0,000	0,106	0,024	0,030	0,041	0,062	0,019	0,021
19	77	0,131	0,111	0,133	0,015	0,133	0,134	0,021	0,023
20	75	0,223	0,222	0,213	0,018	0,207	0,200	0,089	0,095
21	3	0,000	0,000	0,022	0,056	0,028	0,043	0,070	0,074
22	6	0,000	0,184	0,026	0,045	0,052	0,079	0,071	0,075
23	8	0,000	0,022	0,037	0,050	0,108	0,162	0,063	0,067
24	9	0,000	0,000	0,029	0,042	0,062	0,093	0,052	0,055
25	10	0,000	0,083	0,023	0,034	0,031	0,046	0,061	0,065
26	6	0,000	0,018	0,020	0,039	0,029	0,044	0,031	0,033
27	32	0,098	0,041	0,105	0,023	0,108	0,114	0,035	0,037
28	23	0,000	0,092	0,024	0,025	0,037	0,055	0,032	0,034
29	25	0,187	0,211	0,173	0,030	0,151	0,134	0,035	0,037
30	71	0,118	0,227	0,210	0,032	0,188	0,169	0,021	0,022
31	31	0,118	0,059	0,123	0,016	0,125	0,128	0,024	0,026
32	50	0,109	0,156	0,113	0,019	0,112	0,113	0,113	0,120
33	2	0,000	0,000	0,024	0,071	0,037	0,055	0,115	0,122
34	2	0,000	0,000	0,026	0,073	0,047	0,070	0,058	0,061
35	8	0,108	0,000	0,113	0,042	0,112	0,114	0,067	0,071
36	7	0,000	0,000	0,030	0,045	0,065	0,098	0,051	0,054
37	9	0,062	0,197	0,069	0,035	0,062	0,063	0,036	0,038
38	17	0,000	0,019	0,019	0,024	0,023	0,034	0,037	0,040
39	24	0,117	0,022	0,124	0,028	0,134	0,142	0,040	0,043
40	20	0,000	0,070	0,028	0,029	0,052	0,078	0,025	0,027
41	50	0,163	0,145	0,160	0,020	0,156	0,153	0,027	0,029
42	38	0,141	0,114	0,139	0,021	0,133	0,130	0,020	0,022
43	76	0,104	0,090	0,112	0,016	0,120	0,128	0,020	0,022
44	73	0,142	0,149	0,142	0,016	0,139	0,137	0,119	0,127
45	2	0,000	0,000	0,027	0,076	0,051	0,076	0,090	0,095
46	3	0,000	0,052	0,021	0,056	0,023	0,035	0,052	0,055
47	10	0,000	0,072	0,024	0,034	0,044	0,066	0,068	0,072
48	7	0,000	0,172	0,029	0,045	0,068	0,102	0,051	0,054
49	10	0,087	0,364	0,095	0,037	0,099	0,105	0,078	0,083
50	5	0,000	0,000	0,027	0,050	0,053	0,080	0,032	0,034
51	23	0,388	0,092	0,053	0,023	0,056	0,066	0,037	0,039
52	21	0,243	0,195	0,223	0,037	0,198	0,176	0,030	0,032
53	31	0,114	0,184	0,120	0,022	0,121	0,124	0,040	0,042

$$\frac{\lambda}{\lambda + \lambda_2} \sum_{j=1}^J \left[m_j(\hat{b})(1 - m_j(\hat{b})) - (1 - 2m_j(\hat{b}))(m_j(\hat{b})) \right]$$

$$\begin{pmatrix} (1 - m_j(\hat{b})) \frac{1}{2} \Sigma^{-1}(\hat{b}) \\ \vdots \\ (1 - m_j(\hat{b})) \frac{1}{2} \Sigma^{-1}(\hat{b}) K^d(\hat{b}) \end{pmatrix} \left[m_j^2(\hat{b})(1 - m_j(\hat{b}))^2 x_j^T \Sigma^{-1}(\hat{b})(x_j) \right]$$

$$+ \frac{\lambda_2}{\lambda + \lambda_2} \left[\sum_{j=1}^J w_j m_j(\hat{b})(1 - m_j(\hat{b}))(x_j) \right]^T$$

$$\times \left[\sum_{j=1}^J w_j m_j(\hat{b})(1 - m_j(\hat{b}))(x_j) \right]. \quad (4.6)$$

Les preuves de ces théorèmes sont présentées en annexe.

À la section suivante, nous appliquerons ces résultats pour

trouver des estimations approximatives des EQM des estimateurs EB. Cependant, avant cela, il convient de souligner

le point suivant.

Si nous désignons le coefficient de $\lambda/(1 + \lambda_2)^2$ par $B_j(\hat{b})$

et le coefficient de $\lambda_2/(1 + \lambda_2)^2$ par $C_j(\hat{b})$ dans le théorème

2, alors en remarquant que $B_j(\hat{b}) = O(1)$ et $C_j(\hat{b}) = O(K^{-1})$

pour K grand, $EQM(\hat{u})$ est maximisée à $\lambda = (B_j(\hat{b}))/$

$(B_j(\hat{b}) - 2C_j(\hat{b}))$, qui est habituellement très proche de 1. La

distribution a priori résultante avec λ remplaçant λ est la

distribution a priori approximative adaptable aux données le

moins favorable. Dans l'exemple que nous considérons, ce

λ estimé s'avère être égal à 1,003, ce qui confirme

l'observation qui précède.

5. Estimations pour petits domaines pour les Asiatiques

Nous commençons par décrire comment les petits

domaines sont constitués. Considérons le quadruplet

(k_1, k_2, k_3, k_4) , où $k_1 = 1, 2, 3$ ou 4 selon que la personne

est chinoise, philippine, indienne d'Asie ou insulaire. Puis,

$k_2 = 1$ ou 2, selon que la personne est un homme ou une

femme. Puis, $k_3 = 1, 2$ ou 3, selon que la personne

appartient au groupe des 0 à 17 ans, des 18 à 64 ans ou des

65 ans et plus. Enfin, $k_4 = 1, 2, 3$ ou 4, selon que la

personne appartient à une région statistique métropolitaine

(RSM) de taille [499 999, 500 000-999 999, 1 000 000-

2 499 999 ou $\geq 2 500 000$. Un petit domaine est maintenant

numéroté par la formule $24(k_1 - 1) + 12(k_2 - 1) + 4(k_3 - 1) +$

k_4 correspondant au quadruplet (k_1, k_2, k_3, k_4) . Par

exemple, le petit domaine constitué de femmes philippines

de 18 à 64 ans vivant dans une RSM de taille 500 000 à

999 999 porte le numéro 42.

très proches.

Les données de base sont $y_{ij} = 1$ ou 0 si la j^e personne dans le i^e petit domaine ne possède pas (possède) une assurance-maladie ;

$w_{ij} =$ le poids d'échantillonnage appliqué à la j^e unité dans le i^e petit domaine ;

$w_{ij} = w_{ij}/\sum_{j=1}^J w_{ij}$ de sorte que $\sum_{j=1}^J w_{ij} = 1$ pour chaque i ;

$x_{ij1} =$ la taille de la famille de la j^e unité dans le i^e petit domaine ;

$x_{ij2} =$ le niveau d'études de la j^e unité dans le i^e petit domaine ;

$x_{ij3} =$ le revenu familial total de la j^e unité dans le i^e petit domaine ;

Soit $p_{ij} = E(y_{ij})$. Pour l'analyse HB, nous modélisons $\theta_{ij} = \logit(p_{ij}) = b_0 + b_1 x_{ij1} + b_2 x_{ij2} + b_3 x_{ij3} + n_{ij}$

$$j = 1, \dots, n_i; i = 1, \dots, 96.$$

Les estimations par domaine directes sont données par

$\hat{p}_{hw} = \sum_{j=1}^J w_{ij} y_{ij}$. Les estimations hiérarchiques bayésiennes

correspondantes sont données par $\hat{p}_{HB} = \sum_{j=1}^J w_{ij} E(p_{ij} | y)$.

Nous utilisons la méthode MCMC décrite à la section 2 pour

obtenir ces estimations. Dans le tableau elles sont appelées

HB. Les erreurs-types a posteriori connexes sont appelées

é-t. (HB). Dans notre hyperprior, nous considérons : $c =$

0, 2, 0, 02, 0, 002 ; $d = 0, 2, 0, 02, 0, 002$. Les résultats sont

très insensibles au choix des hyperpriors et ne sont présentés

que pour $c = d = 0, 02$. En outre, nous avons des estimateurs

EB pour différents choix du paramètre λ . Les résultats sont

présentés pour $\lambda = 0, 1, 0, 5$ et 1.

Le tableau 1 donne les diverses estimations des propor-

tions d'Asiatiques non assurés et les erreurs-types connexes

pour les divers petits domaines pour l'année 2000. Le do-

maine 2 est exclu parce que la taille d'échantillon est nulle. Il

correspond aux hommes philippines de 0 à 17 ans appartenant

à une RSM de taille de 500 000 à 999 999. Les mesures de

précision (é-t. a posteriori) associées aux estimations HB

sont désignées par é-t. (HB) et sont données par la formule

é-t. 2 (HB) = $\text{var}(\sum_{j=1}^J w_{ij} p_{ij} | y)$. Un des avantages des

estimations HB ou EB est que, pour les domaines dont la

taille d'échantillon est très petite, les estimations directes de

la proportion de personnes non assurées sont souvent nulles,

tandis que les estimations bayésiennes sont faibles mais non

nulles. Nous avons choisi de ne pas regrouper les estimations

directes pour les domaines dont les tailles d'échantillon sont

très faibles. Les covariables au niveau de l'unité étaient assez

différentes et il n'y avait aucun moyen significatif de les

combiner. Nous notons aussi que, quand $\lambda = 0, 5$, c'est-à-

dire quand les estimations directes et synthétiques ont un

ratio de pondération de 1/2, les estimations EB et HB sont

Donc, l'estimateur de Bayes de $\underline{\mu}_{jw} = \sum_{n_j=1}^N w_{nj} \mu_{nj}$ est donné par $\sum_{n_j=1}^N w_{nj} E(\mu_{nj} | Y_{nj})$.

Cependant, en pratique, \mathbf{b} et λ sont inconnus et doivent être estimés d'après les lois marginales des Y_{nj} .

Malheureusement, à part la loi normale, ces lois marginales sont assez compliquées et trouver l'EMV pour les vraisemblances marginales peut devenir une tâche assez redoutable. Nous calculons donc plutôt les estimations basées sur certaines équations d'estimation optimales sans biais (Godambe et Thompson 1989) qui requièrent uniquement l'évaluation des quatre premiers moments de ces lois marginales. Pour cela, nous partons des fonctions d'estimation sans biais élémentaires $ig_{2j} = Y_{nj} - m_{nj}$ et $g_{2j} = (Y_{nj} - m_{nj})^2 - (\lambda + 1)/(\lambda - v_2)A'(m_{nj})$. Afin de construire les équations d'estimation optimales, posons que

$$\mathbf{D}_j^T = \begin{bmatrix} -E\left(\frac{\partial g_{1j}}{\partial \lambda}\right) & -E\left(\frac{\partial g_{1j}}{\partial \mathbf{b}}\right) \\ -E\left(\frac{\partial g_{2j}}{\partial \lambda}\right) & -E\left(\frac{\partial g_{2j}}{\partial \mathbf{b}}\right) \end{bmatrix}.$$

En outre, posons que

$$\Sigma_j = \begin{bmatrix} \mu_{3j} & \mu_{4j} & \mu_{2j} \\ \mu_{4j} & \mu_{4j} & \mu_{2j} \\ \mu_{2j} & \mu_{4j} & \mu_{2j} \end{bmatrix},$$

où $\mu_{rw} = E(Y_{nj} - m_{nj})^r$ est le r^{e} moment central de Y_{nj} basé sur sa loi marginale. Les équations d'estimation optimales

sont alors données par $\sum_{n_j=1}^N \mathbf{D}_j^T \Sigma_j^{-1} \mathbf{g}_{2j} = \mathbf{0}$, où $\mathbf{g}_{2j} = (g_{1j}, g_{2j})^T$. Nous obtenons les estimations de \mathbf{b} et λ (si elles existent) en résolvant ces équations. Nous trouvons les solutions de ces équations à l'aide de l'algorithme de

Nelder-Mead.

Malheureusement, la méthode susmentionnée échoue pour des données binaires. Dans ce cas, $v_2 = -1$, de sorte que $\text{var}(Y_{nj})$ ne dépend pas de λ . En effet, les lois bêta binaires marginales des Y_{nj} sont non identifiables en λ . Un simple moyen de le vérifier est que, si $Y \sim \text{Bin}(1, p)$, et

$p \sim \text{Beta}(\lambda m, \lambda(1 - m))$, alors $E(Y) = E(p) = m$, et une loi binaire est entièrement caractérisé par sa moyenne. Le problème ne se pose pas pour une loi binomiale (n, p) avec $n \geq 2$, puisque, avec la même loi marginale pour p , la fonction génératrice des moments de la loi marginale de la binomiale Y est $E[(p \exp(t) + 1 - p)^n]$ qui dépend de λ .

Pour Y_{nj} binaire, $\partial g_{1j} / \partial \lambda = \partial g_{2j} / \partial \lambda = 0$ de sorte que le deuxième élément du vecteur $\sum_{n_j=1}^N \mathbf{D}_j^T \Sigma_j^{-1} \mathbf{g}_{2j}$ soit nul. Conséquemment, l'approche des équations d'estimation proposées ne permet pas d'estimer λ . Les données de base considérées dans notre application sont binaires, ce qui nécessite la modification de la procédure proposée.

Nous avons donc considéré la fonction d'estimation optimale (pour λ connu)

$$\underline{\mu}_{jw}^{\text{EB}} = \frac{1}{\lambda} Y_{nj} + \frac{\lambda + 1}{\lambda} m_{nj}(\mathbf{b}). \quad (4.4)$$

Par conséquent, l'estimateur EB de $\underline{\mu}_{jw}$ est $\sum_{n_j=1}^N w_{nj} \underline{\mu}_{jw}^{\text{EB}}$.

La méthode décrite ci-dessus repose sur l'hypothèse que λ est connu. Nous pouvons trouver des estimations de μ_{nj} pour divers choix de λ . Dans le présent article, nous avons essayé $\lambda = 0, 1, 0, 5$ et 1, et nous avons comparé les estimations ainsi obtenues aux estimations HB correspondantes.

Ensuite, à la présente section, nous trouvons les erreurs quadratiques moyennes (EQM), ainsi que les EQM estimées de $\underline{\mu}_{jw}^{\text{EB}}$ en supposant que λ est connu. Nous énonçons deux théorèmes dans la présente section. Mais avant cela, il faut que nous précisions certains éléments de notation. Soit

$\mathbf{M} = \text{Diag}(m_1, \dots, m_{n_1}, \dots, m_{n_{n_1}})$ et $\Sigma(\mathbf{b}) = \mathbf{X}^T \mathbf{M} (\mathbf{I} - \mathbf{M}) \mathbf{X} = \sum_{n_j=1}^N \sum_{n_j=1}^N m_{nj} (1 - m_{nj}) \mathbf{x}_{nj} \mathbf{x}_{nj}^T$. En outre, soit $n_T = \sum_{k=1}^K n_k$. Nous supposons que $1 \leq n_k \leq C$ pour chaque k , de sorte que $n_T = O(k)$, où O_ϵ désigne l'ordre exact.

Les deux théorèmes sont donnés ci-après.

Théorème 1. Supposons que $\Sigma(\mathbf{b}) = O_\epsilon(k)$, c'est-à-dire que chaque élément de $\Sigma(\mathbf{b})$ a pour borne inférieure une certaine constante C_1 , et pour borne supérieure, une certaine constante C_2 , où $0 < C_1 < C_2 < \infty$. Alors, une expression appropriée pour $\text{EQM}(\underline{\mu}_{jw}^{\text{EB}})$ correcte jusqu'à l'ordre $O(k^{-1})$ est donnée par

$$\text{EQM}(\underline{\mu}_{jw}^{\text{EB}}) = \frac{\lambda + 1}{\lambda} \sum_{n_j=1}^N w_{nj}^2 m_{nj}(\mathbf{b}) (1 - m_{nj}(\mathbf{b}))$$

$$+ \frac{\lambda + 1}{\lambda^2} \sum_{n_j=1}^N w_{nj}^2 m_{nj}(\mathbf{b}) (1 - m_{nj}(\mathbf{b})) \mathbf{x}_{nj} \mathbf{x}_{nj}^T$$

$$\times \Sigma^{-1}(\mathbf{b}) \left[\sum_{n_j=1}^N w_{nj} m_{nj}(\mathbf{b}) (1 - m_{nj}(\mathbf{b})) \mathbf{x}_{nj} \right]. \quad (4.5)$$

Théorème 2. Supposons que $\Sigma(\mathbf{b}) = O_\epsilon(k)$. Alors, l'approximation suivante de $\text{EQM}(\underline{\mu}_{jw}^{\text{EB}})$ est correcte jusqu'à l'ordre $O(k^{-1})$.

De nouveau, désignons par y_j^* la réponse de la j^e unité dans le i^e petit domaine ($j = 1, \dots, n_j; i = 1, \dots, k$). En outre, nous posons que le modèle de famille exponentielle pour les y_j^* est donné par (3.1), mais nous supposons aussi que les y_j^* possèdent une fonction de probabilité ou une densité de probabilité appartenant à la classe des familles

4. Estimation empirique bayésienne

variation de grande taille.

biases sont assortis d'erreurs-types et de coefficients de tailles d'échantillon sont si petites que ces estimateurs sans mentionner plus haut, pour nombre de ces domaines, les $\bar{y}^{nw} = \sum_{j=1}^n w_j y_j^*$. Cependant, comme nous l'avons l'estimateur sans biais direct de $\bar{\mu}^{nw}$ est donné par $A(\bar{\mu}^{nw} | Y) + \sum_{j=1}^n w_j \text{Cov}(\bar{\mu}^{nw}, \mu_j^* | Y)$. Par contre, $E[\bar{\mu}^{nw} | Y] = \sum_{j=1}^n w_j E(\mu_j^* | Y) = \sum_{j=1}^n w_j^2 E[\bar{\mu}^{nw} | Y] + \sum_{j=1}^n w_j \text{Cov}(\bar{\mu}^{nw}, \mu_j^* | Y)$. D'après ces calculs, nous trouvons maintenant immédiatement $R^{-1} \sum_{r=1}^R (\bar{\mu}^{nw(r)})^2 - (R^{-1} \sum_{r=1}^R \bar{\mu}^{nw(r)})^2 = (R^{-1} \sum_{r=1}^R \bar{\mu}^{nw(r)})^2 - (R^{-1} \sum_{r=1}^R \bar{\mu}^{nw(r)})^2$. Enfin, l'estimation de Monte-Carlo de $\text{Cov}(\bar{\mu}^{nw}, \mu_j^* | Y)$ est donnée par $R^{-1} \sum_{r=1}^R (\bar{\mu}^{nw(r)} \mu_j^{*(r)})^2 - (R^{-1} \sum_{r=1}^R \bar{\mu}^{nw(r)})^2$. De même, l'estimation de Monte-Carlo de $\text{var}(\bar{\mu}^{nw} | Y)$ est $R^{-1} \sum_{r=1}^R (\bar{\mu}^{nw(r)})^2 - (R^{-1} \sum_{r=1}^R \bar{\mu}^{nw(r)})^2$. L'estimation de Monte-Carlo de $E(\mu_j^* | Y)$ est $R^{-1} \sum_{r=1}^R \mu_j^{*(r)}$. Si $\bar{\mu}^{nw}$ désigne la valeur échantillonnée de $\bar{\mu}^{nw}$ produite à partir du r^e tirage, et que le nombre de tirages est R , alors

Tanner (1996):

de cet algorithme, nous renvoyons de nouveau le lecteur à l'algorithme de Métropolis-Hastings. Pour une discussion mais elle ne l'est pas pour θ_j^* , et requiert l'utilisation de σ_j^2 et b est standard, à partir des lois conditionnelles de σ_j^2 et b formules pour le cas binaire. La production d'échantillons d'échantillons à partir des conditionnelles susmentionnées, Note analyse des données est fondée sur la production

$$\begin{aligned} \theta_j^* | b, \sigma_j^2, y &\sim f(y_j^* | \theta_j^*) \exp \left[-\frac{1}{2\sigma_j^2} (\theta_j^* - y_j^*)^2 \right] \\ b | \theta_j^*, \sigma_j^2, y &\sim N((X^T X)^{-1} X^T y, \sigma_j^2 (X^T X)^{-1}) \\ \sigma_j^2 | \theta_j^*, b, y &\sim \text{IG} \left(\sum_{i=1}^n (\theta_j^* - x_i^* b)^2 + c, \frac{2}{k+d} \right) \end{aligned}$$

sont données par

Pour exécuter l'échantillonneur de Gibbs, nous devons trouver les conditionnelles complètes de θ_j^* , b et σ_j^2 . Elles

référer pratique est Tanner (1996, chapitre 6).

général est exposée dans de nombreuses publications. Une employons l'échantillonneur de Gibbs. La méthode MCMC Carlo par chaînes de Markov (MCMC). En particulier, nous utilisons la méthode d'intégration numérique de Monte-Carlo par chaînes de Markov (MCMC). Au lieu de cela, nous pas exécutable de manière analytique. Il s'agit d'une analyse bayésienne non conjuguée qui n'est

exponentielles naturelles à fonctions de variance quadratique ($FEN-FVQ$). Rappelons que $\mu_j^* = E(y_j^* | \theta_j^*) = \psi(\theta_j^*)$. Avec la structure de fonction de variance quadratique, $\text{Var}(y_j^* | \theta_j^*) = \bar{Q}(\mu_j^*) = v_0 + v_1 \mu_j^* + v_2 \mu_j^{*2}$ où v_0, v_1 et v_2 ne sont pas simultanément nuls. Morris (1982, 1983) a caractérisé les lois appartenant à la famille $FEN-FVQ$. Celle-ci comprend les six lois de base, à savoir Bernoulli, ii) Poisson, iii) normale de variance connue, iv) géométrique, v) exponentielle, vi) hyperbolique sécante, ainsi que leurs convolutions. De cette façon, les lois binomiale, binomiale négative et gamma appartiennent aussi à cette famille. Pour la loi de Bernoulli, $v_0 = 0, v_1 = 1$ et $v_2 = -1$. Pour la loi de Poisson, $v_0 = v_2 = 0$ et $v_1 = 1$. Pour la loi normale de variance connue, $\sigma_j^2, \xi_j^2, v_0 = 1$ et $v_1 = v_2 = 0$. De nouveau, nous supposons sans perte de généralité que $\xi_j^2 = 1$. Nous proposons à la présente section des estimateurs EB pour le paramètre canonique du modèle exponentiel. Ensemble, ils constituent une famille $FEN-FVQ$ sur-conjuguée la famille $FEN-FVQ$ de lois ainsi que d'un prior conjugué des moyennes de petit domaine. Pour cela, nous partons de la présente section des estimateurs EB

$$(4.1) \quad \pi(\theta_j^*) = \exp\{\lambda\{m_j^* \theta_j^* - \psi(\theta_j^*)\}\} C(\lambda, m_j^*)$$

conjugué avec la densité de probabilité

Ensemble, ils constituent une famille $FEN-FVQ$ sur-conjuguée la famille $FEN-FVQ$ de lois ainsi que d'un prior conjugué des moyennes de petit domaine. Pour cela, nous partons de la présente section des estimateurs EB

généralité que $\xi_j^2 = 1$.

Ensemble, ils constituent une famille $FEN-FVQ$ sur-conjuguée la famille $FEN-FVQ$ de lois ainsi que d'un prior conjugué des moyennes de petit domaine. Pour cela, nous partons de la présente section des estimateurs EB

généralité que $\xi_j^2 = 1$.

Ensemble, ils constituent une famille $FEN-FVQ$ sur-conjuguée la famille $FEN-FVQ$ de lois ainsi que d'un prior conjugué des moyennes de petit domaine. Pour cela, nous partons de la présente section des estimateurs EB

généralité que $\xi_j^2 = 1$.

Ensemble, ils constituent une famille $FEN-FVQ$ sur-conjuguée la famille $FEN-FVQ$ de lois ainsi que d'un prior conjugué des moyennes de petit domaine. Pour cela, nous partons de la présente section des estimateurs EB

$$(4.2) \quad E(\mu_j^*) = m_j^* ; \text{var}(\mu_j^*) = \bar{Q}(m_j^*) / (\lambda - v_2) ,$$

$$(4.3) \quad \frac{1}{\lambda} y_j^* + \frac{\lambda + 1}{\lambda} m_j^*(b) = \frac{1}{\lambda} y_j^* + \frac{\lambda + 1}{\lambda} m_j^*(b) + \frac{\text{var}(y_j^* | \mu_j^*)}{\text{cov}(y_j^*, \mu_j^*)} (y_j^* - m_j^*(b))$$

D'où le BLUP de μ_j^* est donné par

$$\tilde{Q}(m_j^*) / (\lambda - v_2) = \frac{\lambda + 1}{\lambda} \tilde{Q}(m_j^*)$$

$$E(\mu_j^*) = E(\mu_j^*) = m_j^* ; \text{cov}(y_j^*, \mu_j^*) = \text{var}(\mu_j^*)$$

le montrer, nous calculons

Cette expression peut également être considérée comme le meilleur prédicteur linéaire sans biais (BLUP) de μ_j^* . Pour

$$E(\mu_j^* | y_j^*) = \frac{1}{\lambda} y_j^* + \frac{\lambda + 1}{\lambda} m_j^*(b)$$

est donné par (Morris 1983)

Nous obtenons d'abord l'estimateur de Bayes de μ_j^* . Il ce dernier comme étant le paramètre de précision.

est strictement décroissante en λ , nous pouvons interpréter

$$(4.2) \quad E(\mu_j^*) = m_j^* ; \text{var}(\mu_j^*) = \bar{Q}(m_j^*) / (\lambda - v_2) ,$$

1983),

i^e petit domaine, et g est la fonction lien. Alors (Morris

pour θ_j^* , où $m_j^* = g(x_j^* b)$, $j = 1, \dots, n_j; i = 1, \dots, k$. Ici,

2. Choix des covariables

Comme nous l'avons mentionné dans l'introduction, le nombre de covariables est supérieur à 800. Les inclure toutes dans le modèle initial est impossible et inutile. Nous sommes partis de ce que nous considérons comme un ensemble sensé de six covariables et avons suivi un processus de sélection entièrement séquentiel (avec un seuil de signification de 0,05) pour finalement arriver au meilleur modèle possible.

Les six covariables que nous avons considérées étaient 1) l'état matrimonial légal, 2) la taille de la famille, 3) le niveau d'études, 4) les gains totaux l'année précédente, 5) le revenu familial total et 6) la situation d'emploi à temps plein.

Après la procédure séquentielle, notre modèle final contenait, en plus du terme d'ordonnée à l'origine, la taille de la famille, le niveau d'études et le revenu familial total comme covariables.

Puisque la procédure SURVEYREG de SAS Version 8 ajuste des modèles de régression linéaire et produit des tests d'hypothèse et des estimations pour les données d'enquête, nous l'avons utilisée pour choisir nos covariables. À l'époque où nous avons effectué nos travaux de recherche, la régression logistique pour la sélection des covariables n'était pas disponible. Il convient de souligner que SURVEYREG tient compte de la mise en grappes et de la pondération inégale, et produit des erreurs-types qui tiennent compte correctement des plans de sondage complexes.

3. Analyse hiérarchique bayésienne

Un modèle général de famille exponentielle à un paramètre est donné par

$$f(y_{ij}|\theta_{ij}) = \exp\{\xi_{ij}^T y_{ij} - \psi(\theta_{ij})\} h(y_{ij}; \xi_{ij}), \quad (3.1)$$

$j = 1, \dots, m_i$, $i = 1, \dots, k$. Ici, y_{ij} est la réponse de la j^{e} unité dans le i^{e} petit domaine, tandis que ξ_{ij} les «dits» paramètres de surdispersion, sont supposés connus et pris égaux à 1 sans perte de généralité, parce que l'on peut autrement travailler avec les paramètres transformés, $\xi_{ij} = \xi_{ij}^* \theta_{ij}$. La fonction h est une fonction positive qui dépend des y_{ij} , mais non des θ_{ij} . Si y_{ij} est binaire avec une probabilité de succès p_{ij} alors $\theta_{ij} = \logit(p_{ij})$. Dans notre exemple, y_{ij} la réponse de la j^{e} personne dans le i^{e} petit domaine, est vaut 1 ou 0 selon que la personne ne possède pas ou possède une assurance-maladie. Nous souhaitons estimer $\underline{\mu}^{mw} = \sum_{j=1}^m w_j p_{ij}$ les moyennes de domaine pondérées des proportions de population. Dans ce cas, l'estimateur direct de $\underline{\mu}^{mw}$ est $\sum_{j=1}^m w_j y_{ij}$. Ces estimateurs directs sont habituellement assortis d'erreurs-types et de coefficients de variation importants. Nous supposons que

$$\theta_{ij} = x_{ij}^T \boldsymbol{b} + u_i + n_j, \quad j = 1, \dots, m_i, \quad i = 1, \dots, k, \quad (3.2)$$

L'étape suivante du modèle est section 5.

Nous considérerons des applications particulières à la bjectifive de θ_{ij} . En particulier, $\mu_{ij} = p_{ij}$ dans le cas binaire. Puisque $\psi^*(\theta_{ij}) = \text{var}(y_{ij}|\theta_{ij}) = E(y_{ij}|\theta_{ij}) = \mu_{ij}$ général quand nous voulons estimer $\mu_{ij} = E(y_{ij}|\theta_{ij})$ d'exécuter l'analyse pour le modèle hiérarchique bayésien À la présente section, nous discutons de la façon du domaine n'est pas possible dans cette situation.

des dénombrements en se servant de covariables au niveau individuel et non au niveau du domaine. La modélisation exemple, nous possédons des covariables au niveau domaine. La difficulté tient au fait que, dans le présent dénombrements de personnes non assurées au niveau du utiliser ici un modèle de Poisson fondé sur les Poisson(λ_{ij}), de sorte que $\theta_{ij} = \log(\lambda_{ij})$. Nous pouvons spécifiquement dans le présent article, est $y_{ij} \sim$ Un autre exemple important, qui n'est pas considéré ces poids et nous devons supposer que ceux-ci sont connus. Cependant, les utilisateurs réels utilisés pour produire tenant compte de la poststratification et de la non-réponse. w_j ne sont souvent que des estimations, par exemple en $i = 1, \dots, k$. Il faut toutefois reconnaître qu'en pratique, les normalisons de manière que $\sum_{j=1}^m w_j = 1$ pour tout les poids de sondage w_j sont connus et nous les

x_{kn_j} et supposons que \boldsymbol{X} est une matrice de plein rang. En outre, soit $\boldsymbol{X}^T = (x_{11}, \dots, x_{1m_1}, \dots, x_{kn_1}, \dots, x_{kn_m})^T$. Soit $\boldsymbol{y} = (y_{11}, \dots, y_{1m_1}, \dots, y_{kn_1}, \dots, y_{kn_m})^T$. Alors, la loi a posteriori conjointe est donnée par

$$\pi(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2 | \boldsymbol{y}) \propto \prod_{i=1}^k \prod_{j=1}^{m_i} f(y_{ij} | \theta_{ij}) \times \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^k \sum_{j=1}^{m_i} (\theta_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{b})^2\right] \times (\sigma^2)^{-D/2-1} \exp\left[-\frac{2\sigma^2}{c}\right]. \quad (3.3)$$

Bayes $n^{-1}(1-B_0)+(B-B_0)^2(\bar{y}-\mu)^2$, $B_0=(1+A_0)^{-1}$, qui peut être assez bien supérieur au risque de Bayes correspondant de $\sum_{i=1}^n w_i y_i$, selon, évidemment, les valeurs de B_0, μ et les w_i .

Le présent article décrit la production des estimations par domaine de la proportion de personnes non couvertes par une assurance pour la population asiatique. Les estimations et les mesures de précision sont basées sur le modèle hiérarchique bayésien, ainsi que sur le modèle empirique bayésien. L'analyse a été effectuée pour chaque année de 1997 à 2000, mais par souci de concision, les résultats ne sont présentés que pour l'année 2000. Nous avons exécuté une analyse semblable pour la population hispanique. Dans ce cas, le nombre de petits domaines était égal à 336. Puisque la méthodologie était la même que pour la population asiatique, faute d'espace, nous n'incluons pas cette analyse dans l'article non plus.

Le groupe asiatique est formellement composé 1) des Chinois, 2) des Philippins, 3) des Indiens d'Asie et 4) d'autres, tels que les Coréens, les Vietnamiens, les Japonais, les Hawaïens, les Samoans, les Guyanais, etc. Ces personnes sont affectées à des domaines particuliers selon leur âge, leur race, leur sexe et la région d'où elles proviennent. Il existe trois groupes d'âge (de 0 à 17 ans, de 18 à 64 ans et 65 ans et plus), deux sexes, quatre races et quatre régions selon la taille de la région statistique métropolitaine ($> 2 500\ 000$). Donc, le nombre total de domaines est égal à $3 \times 2 \times 4 \times 4 = 96$. Quand les personnes sont affectées à leurs domaines respectifs, de nombreux domaines ne contiennent que quelques observations. En effet, dans plusieurs d'entre eux, la taille de l'échantillon est égale à 1, tandis que dans un autre, elle est nulle.

Le plan des autres sections est le suivant. À la section 2, nous abordons la sélection des covariables pour la population asiatique. À la section 3, nous discutons de la méthodologie HB générale nécessaire pour obtenir les estimations pour petits domaines et les mesures de précision connexes. À la section 4, nous discutons de l'adéquation du modèle HB proposé. À la section 5, nous discutons d'une méthode EB alternative et nous trouvons des erreurs quadratiques moyennes (EQM) correctes jusqu'à l'ordre deux (à préciser plus loin) des estimateurs EB proposés, ainsi qu'une approximation correcte jusqu'à l'ordre deux de ces EQM. À la section 6, nous trouvons les estimations pour petits domaines et les mesures de précision correspondantes pour la sous-population asiatique en 2000 en utilisant les méthodes HB et EB, et nous les comparons aux estimations directes. Enfin, à la section 7, nous tirons certaines conclusions.

analogue HB du modèle linéaire mixte généralisé (GLMM pour *generalized linear mixed model*) pour obtenir les moyennes et les erreurs-types a posteriori des proportions pour petits domaines de population. La méthode a été proposée par Ghosh, Natarajan, Stroud et Carlin (1998). L'approche EB est fondée sur la théorie des fonctions d'estimation optimales. Nous obtenons les estimateurs EB et les estimateurs approximatifs de l'erreur quadratique moyenne correspondant par une méthode asymptotique analogue à celles de Prasad et Rao (1990) et de Ghosh et Maiti (2004). Alors que celle de Ghosh et Maiti (2004) est fondée sur des données au niveau de la région, la présente approche s'appuie sur des données au niveau de l'unité. Par conséquent, nous devons apporter certaines modifications à la procédure de Ghosh et Maiti (2004) pour élaborer les estimateurs. En outre, comme celle de ces auteurs, la méthodologie générale n'est pas limitée aux données binaires. Elle est applicable à la famille exponentielle naturelle à fonctions de variance quadratiques (Morris 1982, 1983). Le calcul des erreurs quadratiques moyennes des estimations sous le modèle proposé est un peu plus simple que celui de Ghosh et Maiti (2004) pour le cas binaire. En outre, comme ces auteurs, nous utilisons dans notre analyse les poids de sondage ainsi que le modèle pour calculer les estimations pour petits domaines. Donc, dans un certain sens, notre méthode peut être considérée comme une estimation fondée sur un modèle et assistée par le plan de

Les poids de sondage liés aux unités d'échantillonnage des probabilités de sélection. Ils sont souvent utilisés pour produire des estimations sans biais par rapport au plan. L'exemple classique est le fameux estimateur d'Horvitz-Thompson. Cependant, alors que les estimateurs de ce genre protègent contre l'échec du modèle, ils peuvent entraîner une perte d'efficacité si le modèle hypothétique est vérifié. Par exemple, dans des conditions bayésiennes simples, si les y_i/θ_i sont indépendamment distribués de loi $N(\theta_i, 1)$, tandis que les θ_i sont indépendants et identiquement distribués de loi $N(\mu, A)$, $i = 1, \dots, n$, alors l'estimateur de Bayes (moyenne a posteriori) de $\theta = n^{-1} \sum_{i=1}^n \theta_i$ est $n^{-1} \sum_{i=1}^n [(1-B)y_i + B\mu] = (1-B)\bar{y} + B\mu$, où $B = (1+A)^{-1}$. Cet estimateur possède le risque de Bayes $n^{-1}(1-B)$ sous le modèle supposé. Par ailleurs, l'estimateur $\sum_{i=1}^n w_i y_i$ de θ , avec $\sum_{i=1}^n w_i = 1$ possède le risque de Bayes $n^{-1}(1-B) + E[(1-B)\bar{y} + B\mu - \sum_{i=1}^n w_i y_i]^2$. Cependant, si le modèle supposé n'est pas vérifié, par exemple, si les θ_i sont indépendants et identiquement distribués de loi $N(\mu, A)$, où A s'écarte fortement de A_0 , alors l'estimateur $(1-B)\bar{y} + B\mu$ de θ possède un risque de

Estimations pour petits domaines bayésiennes hiérarchiques et empiriques de la proportion de personnes sans assurance-maladie dans les groupes de population minoritaires

Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff et Van L. Parsons¹

Résumé

Le présent article traite de l'estimation pour petits domaines de la proportion de personnes sans assurance-maladie dans divers groupes minoritaires. Les petits domaines sont définis par le croisement de l'âge, du sexe et d'autres caractéristiques démographiques. Des méthodes d'estimation bayésiennes hiérarchiques ainsi qu'empiriques sont appliquées. En outre, des approximations exactes jusqu'à l'ordre deux des erreurs quadratiques moyennes des estimateurs bayésiens empiriques et des estimateurs corrigés du biais de ces erreurs quadratiques moyennes sont fournies. La méthodologie générale est illustrée au moyen d'estimations de la proportion de personnes non assurées pour plusieurs petits domaines de la sous-population asiatique.

Mots-clés : Asiatique ; corrigé du biais ; erreur quadratique moyenne ; exact jusqu'à l'ordre deux.

1. Introduction

La motivation principale de l'étude était l'estimation pour petits domaines de la proportion de personnes sans assurance-maladie dans divers groupes de population minoritaires. Les petits domaines sont constitués par croisement de l'âge, du sexe, de la race et de la région auxquels les personnes appartiennent. Les données de la National Health Interview Survey (NHIS) fournissent la réponse binaire au niveau de l'individu (c'est-à-dire possède ou non une assurance-maladie), ainsi que les covariables à ce niveau. Les données peuvent être obtenues à l'adresse <http://www.cdc.gov/nchs/nhis.htm>. La conception de la NHIS est décrite dans Botman, Moore, Mortariy et Parsons (2000). Durant une année typique, on sélectionne pour la NHIS un échantillon d'unités de logement. L'ensemble des membres de chaque unité est considéré comme un ménage, et ceux entre lesquels existe une « forte » relation sont considérés comme une famille. (Les unités structurelles sont définies de manière plus explicite au chapitre 5.2 du document sur le recensement à l'adresse www.census.gov/prod/2002pubs/ip63rv.pdf). Chaque année, les données de la NHIS sont recueillies auprès d'environ 40 000 ménages, dont plus de 98 % sont des ménages unifamiliaux, ce qui représente environ 100 000 personnes. Pour les questions de type « familial », par exemple celles sur la couverture par une assurance, tous les adultes à domicile sont invités à participer à l'interview, mais la réponse par personne interposée est également permise. Pour les enfants, les réponses sont fournies par un adulte.

Pour toute année de référence, l'enquête originale contient des données sur plus de 100 000 personnes et sur

Pour une sous-population minoritaire visée, la taille d'échantillon dans les domaines n'est pas toujours très grande. Donc, les estimations directes risquent de ne pas être très fiables, et d'être assorties de grandes erreurs-types et de grands coefficients de variation. Il faut alors recourir à des méthodes d'estimation pour petits domaines en se produisant des estimations indirectes de ces domaines en se basant sur des modèles implicites ou explicites. Ces modèles établissent un lien entre les domaines et produisent donc habituellement des estimations par domaine d'une plus grande précision par emprunt de données à d'autres domaines.

Nous nous servons de méthodes hiérarchiques bayésiennes (HB) et empiriques bayésiennes (EB) pour obtenir les estimations pour petits domaines, ainsi que les mesures de précision connexes. L'analyse est fondée sur un

plus de 800 variables. Pour ces personnes, nous possédons des renseignements sur la variable de réponse principale, c'est-à-dire le fait qu'une personne possède une assurance-maladie ou non. En plus, nous disposons d'information sur les caractéristiques démographiques, telles que l'âge, le sexe, la race, la région, le niveau d'études, la catégorie de revenu, les problèmes de santé, les incapacités (s'il en existe) et de nombreux autres facteurs socioéconomiques.

Pour l'ensemble de la population américaine, les estimations directes pour ces domaines, c'est-à-dire les proportions d'échantillon pondérées, sont assez fiables, parce que la taille d'échantillon est raisonnablement grande pour chaque domaine. Par contre, cela n'est pas nécessairement le cas quand l'analyse est axée sur des sous-populations particulières, tels que les Hispaniques, les Asiatiques et des groupes minoritaires similaires de la collectivité.

Bibliographie

- Lavarakas, P. J. (1993). *Telephone Survey Method: Sampling, selection, and supervision*. 2^{ème} Edition. Newbury Park, Calif. : Sage.
- Little, J. A., et Rubin, D. B. (2002). *Statistical analysis with missing data*. 2^{ème} Edition. New York : John Wiley & Sons, Inc.
- Martin, E. A., Traugott, M. W. et Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *The Public Opinion Quarterly*, 69, 342-369.
- Miehleis, B., et Moelenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom drop-out. *Communications in Statistics: Theory and Methods*, 26, 65-94.
- Moelenberghs, G., Kenward, M. G. et Goetghebuer, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Applied Statistics*, 50, 15-29.
- Monterola, C., Lim, M., Garcia, F. et Saloma, C. (2001). Feasibility of a neural network as classifier of undecided respondents in a public opinion survey. *International Journal of Public Opinion Research*, 14, 222-299.
- Myers, D. J., et O'Connor, R. E. (1983). The undecided respondents in mandatory voting settings: A Venezuelan exploration. *The Western Political Quarterly*, 36, 420-433.
- Park, T., et Brown, M. B. (1994). Models for categorical data with the American Statistical Association, 89, 44-52.
- Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54, 1579-1690.
- Perry, P. (1979). Certain problem in election survey methodology. *Public Opinion Quarterly*, 43, 312-325.
- Portnoy, R. F. (1994). Telephone sampling in epidemiologic research: To reap the benefits, avoid the pitfalls. *American Journal of Epidemiology*, 139, 967-978.
- Rubin, D. B., Stern, H. S. et Vehovar, V. (1995). Handling "Don't Know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.
- Smith, P. W. F., Skinner, C. J. et Clarke, P. S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intention data. *Applied Statistics*, 48, 563-577.
- Smith, T. W. (1984). Non attitudes: A review and evaluation. Dans *Surveying Subjective Phenomena*, (Eds. C. F. Turner et E. Martin). New York : Russell Sage Foundation, 2, 215-255.
- Stasny, E. A. (1986). Estimating gross flow using panel data with nonresponse: An example from the Canadian Labor Force survey. *Journal of the American Statistical Association*, 81, 42-47.
- Stasny, E. A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating Gross Labor-Force flows. *Journal of Business and Economic Statistics*, 6, 207-219.
- Lau, R. R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential elections. *Public Opinion Quarterly*, 59, 589-605.
- Kim, T. (1995). Discriminant analysis as a prediction tool for uncommitted voters in pre-election polls. *International Journal of Public Opinion Research*, 7, 110-127.
- Groves, R. M., et Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.
- Gelman, A., Carlin, J. P., Stern, H. S. et Rubin, D. B. (2004). *Bayesian Data Analysis*. 2^{ème} Edition. New York : Chapman and Hall/CRC.
- Forster, J. J., et Smith, R. W. F. (1998). Model-based inference for categorical data subject to non-ignorable non-response. *Journal of the Royal Statistical Society B*, 60, 57-70.
- Fenwick, I., Wiseman, F., Becker, J. F. et Heiman, J. R. (1982). Classifying undecided voters in pre-election polls. *Public Opinion Quarterly*, 46, 383-391.
- Flannelly, K. J., et McLeod, M. S. Jr. (2000). Reducing undecided voters and other sources of error in election surveys. *International Journal of Market Research*, 42, 231-237.
- Dempster, A. P., Laird, N. M. et Rubin, D. M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- De Heer W. (1999). International response trends of an international survey. *Journal of Official Statistics*, 15, 129-142.
- Clogg, C. C., Rubin, D. B., Schenker, N. et Schultz, B. (1991). Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.
- Clarke, P. S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response. *Biometrical Journal*, 44, 701-717.
- Chen, Q. L., et Stasny, E. A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Rapport technique*, Department of Statistics, The Ohio State University.
- Chen, T. (1972). Mixed-up frequencies and missing data in contingency tables. Mémoire de doctorat non publié, University of Chicago. Dept. of Statistics.
- Baker, S. G., Rosenberger, W. F. et Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11, 643-657.
- Baker, S. G., et Laird, N. M. (1988). Regression analysis for categorical variable with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.
- Agresti, A. (2002). *Categorical Data Analysis*. 2^{ème} Edition. New York : John Wiley & Sons, Inc.

Tableau 7
Moyenne des écarts-types et probabilités de couverture à 95 % (entre parenthèses) pour B_{11}^1, R_1^1

	$(\Phi_{11}^{11}, R_{11}^{11}, B_{11}^{11})$	$NON2_{NE}$	$NON2_{SE}$	$NON2_{NE}$	$NON2_{SE}$	$NON2_{NE}$	$NON2_{SE}$	$NON2_{NE}$	$NON2_{SE}$
Solution	(0,2, 0,2)	89,5(0,974)	0,082(0,978)	0,064(0,978)	0,093(0,973)	0,071(0,972)	0,060(0,957)	0,071(0,972)	0,060(0,957)
limite	(0,2, 0,4)	158,3(0,959)	0,072(0,963)	0,096(0,963)	0,079(0,958)	0,066(0,958)	0,058(0,940)	0,066(0,958)	0,058(0,940)
	(0,6, 0,6)	135,3(0,940)	0,065(0,941)	0,057(0,941)	0,071(0,941)	0,062(0,939)	0,056(0,922)	0,062(0,939)	0,056(0,922)
	(0,2, 0,8)	57,4(0,930)	0,070(0,938)	0,061(0,935)	0,076(0,928)	0,066(0,928)	0,060(0,908)	0,066(0,928)	0,060(0,908)
	(0,4, 0,4)	13,3(0,961)	0,079(0,920)	0,061(0,913)	0,069(0,956)	0,072(0,949)	0,060(0,911)	0,072(0,949)	0,060(0,911)
	(0,4, 0,6)	82,2(0,955)	0,072(0,893)	0,059(0,883)	0,086(0,951)	0,069(0,940)	0,058(0,874)	0,069(0,940)	0,058(0,874)
	(0,6, 0,6)	51,2(0,933)	0,071(0,862)	0,059(0,849)	0,084(0,921)	0,068(0,917)	0,059(0,846)	0,068(0,921)	0,059(0,846)
	(0,6, 0,8)	159,6(0,924)	0,071(0,728)	0,057(0,657)	0,089(0,913)	0,069(0,880)	0,058(0,737)	0,069(0,880)	0,058(0,737)
	(0,8, 0,8)	72,8(0,920)	0,070(0,572)	0,056(0,330)	0,093(0,900)	0,070(0,842)	0,058(0,607)	0,070(0,842)	0,058(0,607)
Pas de	(0,2, 0,2)	0,068(0,949)	0,060(0,959)	0,056(0,959)	0,062(0,937)	0,058(0,935)	0,055(0,922)	0,058(0,935)	0,055(0,922)
solution	(0,2, 0,4)	0,066(0,960)	0,060(0,970)	0,056(0,970)	0,061(0,935)	0,058(0,931)	0,055(0,951)	0,058(0,931)	0,055(0,951)
limite	(0,2, 0,6)	0,064(0,940)	0,058(0,945)	0,055(0,945)	0,059(0,959)	0,057(0,919)	0,054(0,909)	0,057(0,919)	0,054(0,909)
	(0,2, 0,8)	0,069(0,933)	0,063(0,944)	0,059(0,941)	0,065(0,926)	0,062(0,925)	0,058(0,920)	0,062(0,925)	0,058(0,920)
	(0,4, 0,4)	0,074(0,910)	0,061(0,836)	0,055(0,828)	0,064(0,899)	0,059(0,884)	0,055(0,824)	0,059(0,884)	0,055(0,824)
	(0,4, 0,6)	0,074(0,915)	0,060(0,815)	0,055(0,806)	0,064(0,922)	0,059(0,879)	0,055(0,792)	0,059(0,879)	0,055(0,792)
	(0,4, 0,8)	0,073(0,891)	0,061(0,786)	0,056(0,771)	0,064(0,873)	0,060(0,852)	0,056(0,763)	0,060(0,852)	0,056(0,763)
	(0,6, 0,6)	0,078(0,859)	0,061(0,567)	0,055(0,470)	0,067(0,853)	0,061(0,795)	0,056(0,572)	0,061(0,795)	0,056(0,572)
	(0,6, 0,8)	0,076(0,843)	0,060(0,515)	0,054(0,402)	0,065(0,817)	0,060(0,767)	0,055(0,556)	0,060(0,767)	0,055(0,556)
	(0,8, 0,8)	0,080(0,755)	0,059(0,110)	0,053(0,017)	0,065(0,728)	0,059(0,607)	0,055(0,158)	0,065(0,728)	0,059(0,607)

5. Conclusion

Nous avons étudié l'application de l'analyse bayésienne à des tableaux de contingence à double entrée incomplets en présence de non-réponse non ignorable. Dans cette situation, les estimations du MV se situent souvent sur la solution limite. Ces solutions limites peuvent donner $G^2 > 0$, même dans le cas d'un modèle saturé (Baker et coll. 1992 ; Park et Brown 1994). Autrement dit, G^2 peut ne pas être appropriée comme statistique pour la spécification des modèles. Pour contourner le problème des solutions limites et obtenir une statistique telle que le facteur de Bayes pour spécifier le modèle indépendamment de l'existence d'une solution limite, nous avons proposé des méthodes d'estimation bayésiennes s'appuyant sur cinq lois a priori différentes. Deux d'entre eux sont nouveaux et les trois autres ont été utilisés antérieurement pour analyser un tableau à simple entrée incomplet. Les deux nouvelles lois entre les répondants et les non-répondants.

L'analyse des données révèle que ces deux nouvelles lois a priori sont plus raisonnables en ce sens qu'ils tiennent mieux compte du mécanisme de non-réponse non ignorable et produisent des estimations proches des résultats réels. En outre, dans le cas des trois lois a priori utilisées antérieurement, notre étude par simulation montre que les estimations

Remerciements

La présente étude a été financée par une subvention de l'Université de la Corée (K0822301).

Nous avons discuté précédemment des questions de pondération à la section 2.2. Cependant, elles requièrent une discussion plus rigoureuse que celle présentée dans cette section. Notre discussion pourrait être approfondie de manière à inclure non seulement divers facteurs de pondération, mais aussi des biais de réponse et d'autres sources de biais et de variations. Ces questions pourront être développées avec soin ultérieurement dans un article élargi.

proches du niveau de couverture nominal.

Nous avons discuté précédemment des questions de pondération à la section 2.2. Cependant, elles requièrent une discussion plus rigoureuse que celle présentée dans cette section. Notre discussion pourrait être approfondie de manière à inclure non seulement divers facteurs de pondération, mais aussi des biais de réponse et d'autres sources de biais et de variations. Ces questions pourront être développées avec soin ultérieurement dans un article élargi.

des estimations du MV pour un tableau de contingence sans solution limite ainsi qu'avec une solution limite, contrairement aux études antérieures. Cependant, quand à lieu une solution limite, les deux nouvelles lois a priori donnent de meilleures résultats que les trois lois a priori antérieures et que les estimations du MV en ce sens qu'ils sont généralement caractérisés par des EQM plus faibles, des biais plus petits et des probabilités de couverture plus proches du niveau de couverture nominal.

Tableau 6
Rapports des EQM moyennes et des biais absolus moyens des estimations bayésiennes relativement à l'estimation du MV quand ne survient aucune solution limite sous un pourcentage de valeurs manquantes de 20 % (Les ratios des biais absolus figurent entre parenthèses)

	$(\beta_{11}^{x_1r_1}, \beta_{11}^{x_2r_2})$	$NON2_{BE}^1$	$NON2_{BE}^2$	$NON2_{BE}^3$	$NON2_{BE}^4$	$NON2_{BE}^5$
Pour	(0,2, 0,2)	0,99(3,37)	1,05(7,00)	0,94(2,51)	0,93(4,89)	1,06(8,96)
	(0,2, 0,4)	0,98(2,57)	1,21(5,13)	0,97(1,89)	1,00(3,26)	1,24(5,56)
	(0,2, 0,6)	1,04(2,18)	1,52(3,84)	0,95(1,67)	1,06(2,38)	1,43(3,71)
	(0,2, 0,8)	1,12(2,04)	1,75(3,53)	1,00(1,48)	1,13(2,14)	1,52(3,21)
	(0,4, 0,4)	1,03(2,40)	1,49(4,66)	0,97(1,69)	1,05(2,74)	1,39(4,46)
$\{m_{j11} + m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$	(0,4, 0,6)	1,20(2,17)	2,11(3,85)	1,00(1,52)	1,22(2,44)	1,78(3,42)
	(0,4, 0,8)	1,28(2,09)	2,36(3,90)	1,05(1,47)	1,25(3,12)	1,86(3,12)
	(0,6, 0,6)	1,22(2,16)	2,49(3,90)	0,96(1,48)	1,21(2,15)	1,90(3,32)
	(0,6, 0,8)	1,52(1,99)	3,19(3,39)	1,11(1,38)	1,45(1,91)	2,29(2,77)
	(0,8, 0,8)	1,66(1,96)	3,64(3,27)	1,14(1,36)	1,52(1,83)	2,43(2,59)
Pour	(0,2, 0,2)	0,88(2,59)	0,89(5,66)	0,87(2,26)	0,89(4,55)	1,21(8,69)
	(0,2, 0,4)	0,93(2,40)	1,27(4,86)	0,93(1,78)	1,00(3,08)	1,50(5,29)
	(0,2, 0,6)	1,09(2,11)	1,93(3,97)	0,98(1,40)	1,15(2,29)	1,85(3,61)
	(0,2, 0,8)	1,24(2,13)	2,36(3,90)	1,02(1,48)	1,27(2,18)	2,06(3,19)
$\{m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$	(0,4, 0,4)	1,03(2,18)	1,81(4,30)	0,96(1,60)	1,12(2,62)	1,85(4,39)
	(0,4, 0,6)	1,23(2,28)	2,62(4,28)	0,99(1,48)	1,29(2,42)	2,28(3,80)
	(0,4, 0,8)	1,42(2,05)	3,26(3,70)	1,07(1,42)	1,44(2,07)	2,53(3,09)
	(0,6, 0,6)	1,33(2,07)	3,22(3,95)	0,99(1,36)	1,36(2,14)	2,54(3,43)
	(0,6, 0,8)	1,65(2,09)	4,14(3,74)	1,13(1,43)	1,61(2,07)	2,98(3,13)
	(0,8, 0,8)	1,91(2,02)	4,48(3,50)	1,16(1,39)	1,66(1,93)	3,03(2,83)

95 % sont les taux de couverture pour les intervalles de confiance à 95 % nominaux. Quand survient une solution limite, même si la probabilité de couverture de l'estimation du MV est celle qui s'approche le plus du niveau de couverture nominal de 95 %, l'estimation du MV possède un écart-type trop grand pour pouvoir être utilisé en pratique. Ces grands écarts-types sont dus au problème des solutions limites de l'estimation du MV. Parmi les estimations bayésiennes, les probabilités de couverture de $NON2_{BE}^3$ sont celles qui s'approchent le plus du niveau de couverture nominal de 95 %, tandis que celles des autres estimations sont généralement inférieures à ce niveau. Cela signifie qu'à part $NON2_{BE}^3$, les estimations bayésiennes sous-estiment les écarts-types.

Quand aucune solution limite ne se présente (deuxième partie du tableau 7), les écarts-types de l'estimation du MV sont nettement plus stables que ceux obtenus dans le cas d'une solution limite. La probabilité de couverture diminue à mesure que $\beta_{11}^{x_1r_1}$ et $\beta_{11}^{x_2r_2}$ augmentent. En particulier, les probabilités de couverture de NON_{BE}^1 , NON_{BE}^2 et NON_{BE}^5 sont considérablement plus faibles que le niveau de couverture nominal de 95 % quand les profils de réponse des répondants et des électeurs indécis sont fort différents (c'est-à-dire $\beta_{11}^{x_1r_1} \geq 0,6$ et $\beta_{11}^{x_2r_2} \geq 0,6$).

Park et Brown (1994) ont utilisé $NON2_{BE}^2$ pour estimer les effectifs attendus par case dans un tableau à simple entrée incomplet sous un modèle de non-réponse non ignorable. Ils ont montré par des études en simulation que $NON2_{BE}^2$ avait une EQM plus petite que l'estimation du MV bien que son biais soit plus important. Cependant, les valeurs supérieures à 1 pour $NON2_{BE}^2$ dans les tableaux 5 et 6 indiquent qu'il n'en est pas ainsi dans un tableau à double entrée incomplet, indépendamment de la solution limite, et que les méthodes bayésiennes ne donnent pas systématiquement de meilleurs résultats que le MV, même en cas d'une solution limite. L'une des raisons pour lesquelles nos résultats de simulation diffèrent de ceux de Park et Brown (1994) quand a lieu une solution limite tient au choix de $(\beta_{11}^{x_1r_1}, \beta_{11}^{x_2r_2})$. Park et Brown n'ayant exécuté leur simulation que dans les conditions où $\beta_{11}^{x_1r_1} = \beta_{11}^{x_2r_2} = 0,34$. Comme le montre le tableau 5, $NON2_{BE}^2$ est meilleure que le MV quand $\beta_{11}^{x_1r_1} \leq 0,4$ et $\beta_{11}^{x_2r_2} \leq 0,4$, tandis que $NON2_{BE}^2$ est pire que le MV quand les profils de réponse des répondants et des non-répondants sont fort différents (c'est-à-dire $\beta_{11}^{x_1r_1} \geq 0,6$ ou $\beta_{11}^{x_2r_2} \geq 0,6$).

Le tableau 7 donne la moyenne des écarts-types et des probabilités de couverture à 95 % pour $\beta_{11}^{x_1r_1}$. Ici, nous avons utilisé la formule de variance donnée dans (9) pour calculer les écarts-types, et les probabilités de couverture à

cinq estimations bayésiennes et de l'estimation du MV

Nous calculons les erreurs quadratiques moyennes (EQM) et les biais absolus de $NON2_{BE}^{AD}$, $NON2_{BE}^{MD}$, $NON2_{BE}^{MD}$ pour $\{\sum_{i=1}^J m_{ij}, i, j = 1, 2\}$. Puis, nous prenons la moyenne sur les quatre EQM et sur les quatre biais absolus que nous obtenons à partir de chaque estimation pour voir la performance globale de l'estimation. De même, nous calculons les EQM moyennes et les biais absolus moyens pour $\{m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$ pour voir la performance de chaque estimation en présence d'imputation des non-réponses.

Le tableau 5 donne les ratios des EQM moyennes et des biais absolus moyens des cinq estimations bayésiennes (c'est-à-dire $NON2_{BE}^{MD}$, $NON2_{BE}^{MD}$, $NON2_{BE}^{MD}$) par rapport à l'estimation du MV (c'est-à-dire $NON2_{BE}^{MD}$) quand survieusement des solutions limites, tandis que le tableau 6 donne les mêmes ratios quand aucune solution limite ne se présente. Donc, les valeurs inférieures à 1 impliquent que l'estimation bayésienne correspondante possède une EQM moyenne ou un biais absolu moyen plus faible que l'estimation du MV. Les deux tableaux ne présentent que les cas pour $\beta_{11}^{x_1 r_1} > \beta_{11}^{x_2 r_2}$ et pour un pourcentage de valeurs

Tableau 5
Ratios des EQM moyennes et des biais absolus moyens des estimations bayésiennes relativement à l'estimation du MV quand survieusement des solutions limites sous un pourcentage de valeurs manquantes de 20 % (les ratios des biais absolus figurent entre parenthèses)

	$(\beta_{11}^{x_1 r_1}, \beta_{11}^{x_2 r_2})$	$NON2_{BE}^1$	$NON2_{BE}^2$	$NON2_{BE}^3$	$NON2_{BE}^4$	$NON2_{BE}^5$
Pour $\{m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$	(0,2, 0,2)	0,68(0,66)	0,47(0,22)	0,76(0,76)	0,65(0,48)	0,42(0,05)
	(0,2, 0,4)	0,68(0,48)	0,57(0,20)	0,77(0,68)	0,60(0,30)	0,56(0,30)
	(0,2, 0,6)	0,67(0,23)	0,73(0,66)	0,77(0,57)	0,64(0,10)	0,69(0,64)
	(0,2, 0,8)	0,77(0,26)	1,08(1,55)	0,83(0,43)	0,76(0,28)	0,95(1,34)
	(0,4, 0,4)	0,65(0,32)	0,69(0,57)	0,76(0,63)	0,61(0,17)	0,65(0,52)
	(0,4, 0,6)	0,58(0,14)	0,83(0,90)	0,71(0,56)	0,56(0,06)	0,69(0,71)
	(0,4, 0,8)	0,75(0,36)	1,46(2,07)	0,78(0,42)	0,74(0,61)	1,12(1,61)
	(0,6, 0,6)	0,66(0,22)	1,35(1,73)	0,73(0,43)	0,66(0,16)	1,01(1,29)
	(0,6, 0,8)	0,85(0,87)	2,27(3,19)	0,76(0,17)	0,83(0,81)	1,52(2,35)
	(0,8, 0,8)	1,12(1,93)	3,58(5,49)	0,83(0,24)	1,04(1,67)	2,18(3,95)
Pour $\{m_{j12} + m_{j21} + m_{j22}, i, j = 1, 2\}$	(0,2, 0,2)	0,57(0,63)	0,27(0,13)	0,69(0,74)	0,41(0,40)	0,28(0,31)
	(0,2, 0,4)	0,54(0,46)	0,37(0,34)	0,68(0,68)	0,42(0,24)	0,44(0,57)
	(0,2, 0,6)	0,51(0,19)	0,69(0,94)	0,65(0,55)	0,47(0,10)	0,69(0,88)
	(0,2, 0,8)	0,63(0,35)	1,39(2,08)	0,71(0,34)	0,62(0,47)	1,11(1,52)
	(0,4, 0,4)	0,49(0,35)	0,54(0,64)	0,65(0,64)	0,42(0,17)	0,57(0,76)
	(0,4, 0,6)	0,48(0,17)	0,98(1,24)	0,62(0,51)	0,45(0,17)	0,85(1,04)
	(0,4, 0,8)	0,62(0,44)	1,81(2,33)	0,67(0,35)	0,61(0,55)	1,35(1,81)
	(0,6, 0,6)	0,55(0,42)	1,70(1,90)	0,63(0,41)	0,54(0,40)	1,28(1,51)
	(0,6, 0,8)	0,78(0,92)	2,91(3,43)	0,69(0,14)	0,75(0,92)	1,96(2,64)
	(0,8, 0,8)	1,13(1,96)	4,63(5,72)	0,75(0,33)	1,02(1,77)	2,86(4,24)

manquantes de 20 %, parce que les EQM et les biais sont presque symétriques autour de la coordonnée de $(\beta_{11}^{x_1 r_1}, \beta_{11}^{x_2 r_2})$. Ils augmentent si nous faisons passer le pourcentage de valeurs manquantes à 30 % en maintenant les mêmes profils d'EQM et de biais que pour le cas où 20 % de données manquent.

Le tableau 5, dans lequel survient une solution limite, montre que $NON2_{BE}^1$, $NON2_{BE}^2$, $NON2_{BE}^3$ ont une EQM plus faible que l'estimation du MV (c'est-à-dire $NON2_{BE}^{MD}$) pour toutes les valeurs de $\beta_{11}^{x_1 r_1}$ et $\beta_{11}^{x_2 r_2}$, sauf $(\beta_{11}^{x_1 r_1}, \beta_{11}^{x_2 r_2}) = (0,8, 0,8)$. Ici, $NON2_{BE}^3$ a une plus petite EQM que l'estimation du MV. Cela est vrai pour les biais absolus. Par ailleurs, le tableau 6, où ne survient aucune solution limite, montre que seule $NON2_{BE}^5$ est comparable à l'estimation du MV en ce qui concerne l'EQM, mais qu'elle est légèrement biaisée. En particulier, $NON2_{BE}^3$ a une EQM plus faible que l'estimation du MV à condition que $\beta_{11}^{x_1 r_1} \neq 0,8$ ou $\beta_{11}^{x_2 r_2} \neq 0,8$ (c'est-à-dire que les profils de réponse des répondants et des non-répondants diffèrent peu).

La grandeur de cette différence peut être mesurée à l'aide des termes les plus importants, $\beta_{11}^{X_1 R_1}$ et $\beta_{11}^{X_2 R_2}$, dans le modèle 2 de non-réponse non ignorable. Puisque

$$\beta_{11}^{X_1 R_1} = \frac{1}{4} \log \frac{m_{1111}^{1121}}{m_{1121}^{1121}} = -0,3254$$

et

$$\beta_{11}^{X_2 R_2} = \frac{1}{4} \log \frac{m_{1111}^{1112}}{m_{1112}^{1112}} = 0,5981, \beta_{11}^{X_1 R_1}$$

Nous considérons un tableau de contingence 2×2 avec cinq estimations bayésiennes décrites à la section 2 pour divers pourcentages de données manquantes et divers profils de réponse sous le modèle de non-réponse non ignorable suivant (c'est-à-dire le modèle 2) :

$$\log(m^{ijkl}) = \beta_0 + \beta_i^{X_1} + \beta_j^{X_2} + \beta_k^{R_1} + \beta_l^{R_2} + \beta_{ik}^{X_1 R_1} + \beta_{jl}^{X_2 R_2} + \beta_{il}^{X_1 X_2} + \beta_{kl}^{R_1 R_2}.$$

Donc, nous comparons uniquement $NON2_{MC}$ et $NON2_{BE}$ pour $i = 1, \dots, 5$ dans cette étude par simulation. Puisque chacune des variables X_1, X_2, R_1 et R_2 comprend deux niveaux, huit paramètres doivent être déterminés pour l'étude par simulation. D'après les équations de

$$4\beta_{11}^{X_1 R_1} = \log \frac{m_{1111}^{1111}}{m_{1111}^{1121}} \text{ et } 4\beta_{11}^{X_2 R_2} = \log \frac{m_{1111}^{1112}}{m_{1112}^{1112}},$$

$$\beta_{11}^{X_1 R_1} = \beta_{11}^{X_2 R_2} = 0$$

signifie qu'il n'y a aucune différence entre les profils de réponses des répondants et des électeurs indécis. Ces profils de réponses diffèrent d'autant plus que les paramètres $\beta_{11}^{X_1 R_1}$ et $\beta_{11}^{X_2 R_2}$ sont grands. Nous faisons varier ces deux paramètres de 0,2 à 0,8 par incréments de 0,2. Nous fixons le pourcentage de valeurs manquantes à 20 % et à 30 % en ajustant $\beta_{11}^{X_1 R_1}$ et $\beta_{11}^{X_2 R_2}$ en posant que

$$\frac{m_{1111}^{1111}/m_{1111}^{1121}}{m_{1111}^{1111}/m_{1112}^{1112}} = 5, \frac{m_{1111}^{1121}/m_{1121}^{1121}}{m_{1111}^{1112}/m_{1112}^{1112}} = 2,$$

et

$$N = \sum_{ijkl} m^{ijkl} = 1\,000.$$

Cela signifie que la taille et le pourcentage de valeurs manquantes pour la case de $X_1 = 1$ et $X_2 = 1$ sont approximativement égaux à cinq fois et deux fois la taille des trois autres cases, respectivement.

Nous produisons un grand nombre d'échantillons $\{Y^{ijkl}, i, j, k, l = 1, 2\}$ dans les conditions susmentionnées jusqu'à ce que nous obtenions 1 000 échantillons aléatoires avec des solutions limitées et 1 000 autres sans solution limite. L'occurrence d'une solution limite est déterminée par le critère donné dans Michiels et Molenberghs (1997) (voir aussi Clarke 2002, ainsi que Smith et coll. 1999 pour plus de précisions). En utilisant $\{Y^{ij11}, Y^{ij12}, Y^{ij21}, Y^{ij22}\}$, $i, j = 1, 2$, obtenu d'après les données générées, nous estimons les effectifs attendus par case m^{ijkl} au moyen des

La grandeur de cette différence peut être mesurée à l'aide des termes les plus importants, $\beta_{11}^{X_1 R_1}$ et $\beta_{11}^{X_2 R_2}$, dans le

plus grand que celui des électeurs indécis à l'égard de Montgomerie, ce qui implique que la plupart des électeurs indécis n'indiquant pas qui est leur candidat préféré voteront vraisemblablement pour Corday comme procureur général. Cela confirme également la croyance populaire selon laquelle les électeurs ont tendance à demeurer « indécis » dans un sondage s'ils appuient un candidat qui est considéré comme étant inférieur dans une course électorale et ils sont enclins à s'abstenir de voter s'ils appuient le candidat qui domine la course avec certitude.

$$\left(\text{c'est-à-dire } \frac{m_{1111}^{1111}/m_{1121}^{1111}}{m_{1112}^{1111}/m_{1121}^{1112}} = e^{4 \times 0,5981} = 10,94 \right)$$

chez les électeurs décidés est environ 10,94 fois vraisemblablement, le taux de soutien pour Montgomerie scrutin. Par ailleurs, parmi les personnes qui voteront stratégie en vue d'accroître la participation des électeurs au qui veut dire que Montgomerie doit mettre en œuvre une plus grande que la possibilité pour les électeurs décidés, ce

$$\left(\text{c'est-à-dire } \frac{m_{1111}^{1111}/m_{1121}^{1111}}{m_{1112}^{1111}/m_{1121}^{1112}} = e^{4 \times -0,3254} = 3,67^{-1} \right)$$

votent pas est d'environ 3,67 fois que les électeurs indécis votent par rapport à celle qu'ils ne parmi les électeurs en faveur de Montgomerie, la possibilité mais qui n'indiquent pas quel est leur candidat préféré. Donc, parmi les électeurs indécis qui voteront vraisemblablement, décidés qui voteront vraisemblablement et le même ratio nombre d'électeurs en faveur de Corday parmi les électeurs ratio du nombre d'électeurs en faveur de Montgomerie au rapport des cotes qui montre la log-différence entre le rendre aux urnes. Par contre, $\beta_{11}^{X_2 R_2}$ est le logarithme du gomery, mais qui n'expriment pas leur probabilité de se même ratio parmi les électeurs indécis qui préfèrent Mont- ment pas » parmi les électeurs décidés pour Montgomerie et le vraisemblablement » à ceux qui « ne voteront vraisemblable- logarithmique entre le ratio du nombre de ceux « qui voteront est le logarithme du rapport des cotes qui montre la différence

$NON2_{ML}$ donne la pire prédiction pour le gouvernement, le maître et le procureur général, parce que les valeurs se situent sur une solution limite; par contre, elle donne la meilleure prédiction pour le trésorier, parce que les valeurs ne se situent pas sur une solution limite. Dans le cas de l'élection du procureur général, $NON2_{BE}$ et $NON2_{BE}^4$ prédisent non seulement le résultat réel exact, mais diffèrent aussi assez bien des autres estimations. Puisque $NON2_{BE}^{9B}$ et $NON2_{BE}^4$ utilisent les lois a priori destinées à refléter les profils de réponse différents des répondants et des électeurs indécis, nous pouvons inférer que la préférence de ces derniers pour un candidat diffère assez fortement de celle des répondants (autrement dit, $NON2_{BE}^3$ et $NON2_{BE}^4$ affectent 19,4 % des électeurs indécis qui voteront vraisemblablement à Montgomey et 80,6 % à Corday, tandis que dans le tableau 2, les données indiquent que le pourcentage allant à Montgomey par opposition à Corday est de 29,4 % contre 70,6 % chez les répondants qui voteront vraisemblablement).

Afin de visualiser cette différence entre les répondants et les électeurs indécis en ce qui a trait aux estimations des paramètres et d'examiner l'effet de l'occurrence de la solution limite sur les estimations sous le modèle 2 de non-estimations du MV et les estimations $NON2_{BE}^3$, ainsi que les écarts-types correspondants pour l'élection du procureur général. Comme il existe une solution limite, toutes les estimations du MV ont un écart-type trop grand comme il fallait s'y attendre. Par ailleurs, $NON2_{BE}^3$ est très stable. Puisque $\beta_{11}^{x_1x_2} = 0,0472$ est la plus petite et que son écart-type est relativement grand, nous négligeons le terme $\beta_{11}^{x_1x_2}$ pour rendre l'interprétation moins complexe. Sous $\beta_{11}^{x_1x_2} = 0$, il n'est pas difficile de montrer qu'en utilisant les estimations de $NON2_{BE}^3$ du tableau 4,

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_1} - \beta_{11}^{x_1r_1}) = 1,3916$$

les personnes qui votent pour Montgomey sont 2,46 fois plus nombreuses que celles qui votent pour Corday parmi les répondants, tandis qu'en vertu de

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_1} - \beta_{11}^{x_1r_1}) = 1,3916,$$

les personnes qui voteront vraisemblablement ($i = 1$) sont 4,02 fois (c'est-à-dire $e^{1,3916}$) plus nombreuses que les personnes qui ne voteront vraisemblablement pas ($i = 2$) parmi les électeurs indécis ($k = 2$), en vertu de

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_1} - \beta_{11}^{x_1r_1}) = 1,3916,$$

les personnes qui voteront vraisemblablement (c'est-à-dire $i = 1$) sont 1,09 fois (c'est-à-dire $e^{0,09}$) plus nombreuses que celles qui ne voteront vraisemblablement pas (c'est-à-dire $i = 2$) parmi les répondants ($k = 1$), tandis qu'en vertu de

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_1} + \beta_{11}^{x_1r_1}) = 0,09,$$

pour chaque valeur fixée de i et k . Donc, en vertu de

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_2} - \beta_{11}^{x_2r_2}) = -1,4942$$

les personnes qui ne voteront vraisemblablement pas sont 4,46 fois plus nombreuses que celles qui voteront vraisemblablement parmi les électeurs indécis. Cela implique que le profil de réponse des répondants et des électeurs indécis est fort différent.

et

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_1} - \beta_{11}^{x_1r_1}) = 1,3916$$

pour chaque valeur fixée de j et l , et

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_1} + \beta_{11}^{x_1r_1}) = 0,09$$

$$\log \frac{m_{12L}}{m_{2L}} = 2(\beta_{11}^{x_2} - \beta_{11}^{x_2r_2}) = -1,4942,$$

Tableau 4
MV et le troisième type d'estimation bayésienne sous le modèle 2 de non-réponse non ignorable pour l'élection du procureur général (les écarts-types sont indiqués entre parenthèses)

β_0	$\beta_{11}^{x_1}$	$\beta_{11}^{x_2}$	$\beta_{11}^{r_1}$	$\beta_{11}^{r_2}$	$\beta_{11}^{x_1r_1}$	$\beta_{11}^{x_2r_2}$	$\beta_{11}^{x_1r_2}$	$\beta_{11}^{r_1r_2}$
$NON2_{ML}$	-3,3735	3,2134	4,8496	4,8186	2,0283	-2,7594	-0,0452	-1,5588
	(3,120)	(8,515)	(3,996)	(8,871)	(3,120)	(8,512)	(0,045)	(2,501)
$NON2_{BE}$	0,6860	0,3704	-0,1490	3,3024	2,2942	-0,3254	0,5981	0,0472
	(0,118)	(0,052)	(2,501)	(2,501)	(0,117)	(0,052)	(0,041)	(2,501)

$NON2_{BE}^3$, $NON2_{BE}^4$ et $NON2_{BE}^5$ dans le premier groupe et les quatre autres estimations, $NON2_{BE}^1$, $NON2_{BE}^2$, IGI_{ML}^2 et $NON3_{ML}^3$, dans le deuxième groupe. Comme il fallait s'y attendre, puisque les lois a priori θ^{ML} pour $NON2_{BE}^3$ et $NON2_{BE}^4$ sont définis de façon telle que l'estimation de m^{ML} diminue et se rapproche du MV sous un modèle de non-réponse ignorable, ces deux estimations bayésiennes sont très proches de IGI_{ML}^2 et offrent par conséquent peu d'avantage par rapport à cette dernière estimation. Il est également intéressant de souligner que $NON3_{ML}^3$ est presque identique à IGI_{ML}^2 , bien que leurs modèles log-linéaires soient spécifiés différemment.

Il n'existe aucun critère général pour déterminer s'il convient d'utiliser un modèle de non-réponse ignorable ou un modèle de non-réponse non ignorable. Cependant, comme l'indique Chen et Stasny (2003), l'hypothèse de non-ignorabilité d'une non-réponse peut être raisonnable dans l'étude fondée sur le sondage électoral de l'État de l'Ohio parce que les gens hésitent à indiquer leur appui pour un candidat impopulaire ou que leurs préférences au moment du sondage ne sont pas fermes ou exactes. À cet égard, les estimations $NON2_{BE}^3$, $NON2_{BE}^4$ et $NON3_{ML}^3$ ne sont peut-être pas appropriées dans ces études de cas particulières, parce qu'elles sont presque les mêmes que les estimations IGI_{ML}^2 du modèle I.

La partie inférieure du tableau donne les prédictions des résultats des élections en utilisant les réponses « votera vraisemblablement » et « ne votera vraisemblablement pas » pour voir ce qui se passerait si les personnes ayant déclaré qu'elles ne « voteraient vraisemblablement pas » se rendaient effectivement aux urnes. La comparaison des deux tableaux nous permet de conclure que les gagnants des élections du gouverneur, du procureur général et du trésorier restent les mêmes quelle que soit la probabilité déclarée de voter, tandis que le gagnant de l'élection du maire aurait pu changer si la plupart des personnes ayant déclaré qu'elles « ne voteraient vraisemblablement pas » avaient effectivement voté.

En nous basant sur le tableau 3, nous pouvons classer les sept estimations, sauf $NON2_{ML}^2$, en deux groupes :

Tableau 3
Prédiction des résultats des élections basée sur les sondages électoraux de l'État de l'Ohio d'octobre 98 et d'avril 98 (les résultats sont exprimés en pourcentage et les chiffres entre parenthèses sont les écarts-types)

	Gouverneur					Maire					Procureur général					Trésorier				
	Fisher	Taft	Autre	Coleman	Teater	Espy	Montgomery	Cordray	Deters	Donofrio	Fisher	Taft	Autre	Coleman	Teater	Espy	Montgomery	Cordray	Deters	Donofrio
$NON2_{ML}^1$	32,2(2,75)	42,1(3,00)	24,8	31,5(4,65)	25,3(4,23)	43,2	75,6(3,71)	24,4	57,0(3,48)	43,0	33,2(2,75)	42,1(3,00)	24,8	31,5(4,65)	25,3(4,23)	43,2	75,6(3,71)	24,4	57,0(3,48)	43,0
$NON2_{BE}^1$	40,6(3,04)	48,5(3,27)	10,9	38,1(5,14)	34,2(4,78)	27,7	72,1(3,61)	27,9	52,7(3,36)	47,3	40,6(3,04)	48,5(3,27)	10,9	38,1(5,14)	34,2(4,78)	27,7	72,1(3,61)	27,9	52,7(3,36)	47,3
$NON2_{BE}^2$	49,3(3,01)	50,7(3,20)	8,40	39,9(5,03)	33,6(4,83)	26,5	71,0(3,59)	29,0	52,1(3,34)	47,9	49,3(3,01)	50,7(3,20)	8,40	39,9(5,03)	33,6(4,83)	26,5	71,0(3,59)	29,0	52,1(3,34)	47,9
$NON2_{BE}^3$	35,8(2,85)	44,5(3,08)	19,7	33,6(4,87)	29,3(4,51)	35,1	63,0(3,67)	37,0	54,3(3,41)	45,7	35,8(2,85)	44,5(3,08)	19,7	33,6(4,87)	29,3(4,51)	35,1	63,0(3,67)	37,0	54,3(3,41)	45,7
$NON2_{BE}^4$	36,3(2,87)	45,2(3,11)	18,6	33,9(4,91)	29,4(4,52)	34,6	63,0(3,64)	37,0	53,9(3,40)	46,1	36,3(2,87)	45,2(3,11)	18,6	33,9(4,91)	29,4(4,52)	34,6	63,0(3,64)	37,0	53,9(3,40)	46,1
$NON2_{BE}^5$	38,9(2,99)	47,4(3,20)	13,7	37,7(4,99)	33,6(4,77)	28,7	66,0(3,54)	34,0	51,5(3,32)	48,5	38,9(2,99)	47,4(3,20)	13,7	37,7(4,99)	33,6(4,77)	28,7	66,0(3,54)	34,0	51,5(3,32)	48,5
IGI_{ML}^2	40,6(3,03)	51,2(3,28)	8,20	40,8(5,16)	33,3(4,76)	25,8	70,9(3,59)	29,1	51,8(3,32)	48,2	40,6(3,03)	51,2(3,28)	8,20	40,8(5,16)	33,3(4,76)	25,8	70,9(3,59)	29,1	51,8(3,32)	48,2
$NON3_{ML}^3$	40,6(3,03)	51,2(3,28)	8,20	40,9(5,16)	33,3(4,75)	25,8	70,9(3,58)	29,1	51,7(3,32)	48,3	40,6(3,03)	51,2(3,28)	8,20	40,9(5,16)	33,3(4,75)	25,8	70,9(3,58)	29,1	51,7(3,32)	48,3
Limite	45	50	5	39	37	24	63	37	57	43	45	50	5	39	37	24	63	37	57	43
Utilisation des réponses « votera vraisemblablement » + « ne votera vraisemblablement pas »																				
$NON2_{ML}^1$	32,7(1,83)	39,4(1,91)	27,8	24,8(2,45)	26,2(2,49)	49,0	77,0(1,64)	23,0	60,2(1,93)	39,8	32,7(1,83)	39,4(1,91)	27,8	24,8(2,45)	26,2(2,49)	49,0	77,0(1,64)	23,0	60,2(1,93)	39,8
$NON2_{BE}^1$	41,3(1,93)	46,4(1,96)	12,3	30,7(2,68)	37,1(2,75)	32,2	72,8(1,74)	27,2	56,0(1,96)	44,0	41,3(1,93)	46,4(1,96)	12,3	30,7(2,68)	37,1(2,75)	32,2	72,8(1,74)	27,2	56,0(1,96)	44,0
$NON2_{BE}^2$	41,9(1,93)	49,2(1,95)	8,90	32,7(2,63)	36,5(2,76)	30,8	71,4(1,77)	28,6	55,3(1,96)	44,7	41,9(1,93)	49,2(1,95)	8,90	32,7(2,63)	36,5(2,76)	30,8	71,4(1,77)	28,6	55,3(1,96)	44,7
$NON2_{BE}^3$	35,4(1,87)	41,8(1,93)	22,7	27,8(2,55)	30,5(2,62)	41,7	61,0(1,72)	39,0	57,6(1,95)	42,4	35,4(1,87)	41,8(1,93)	22,7	27,8(2,55)	30,5(2,62)	41,7	61,0(1,72)	39,0	57,6(1,95)	42,4
$NON2_{BE}^4$	36,0(1,88)	42,6(1,93)	21,4	28,7(2,57)	30,6(2,62)	40,7	60,9(1,75)	39,1	57,2(1,95)	42,8	36,0(1,88)	42,6(1,93)	21,4	28,7(2,57)	30,6(2,62)	40,7	60,9(1,75)	39,1	57,2(1,95)	42,8
$NON2_{BE}^5$	39,1(1,91)	45,1(1,95)	15,8	30,7(2,63)	35,8(2,74)	33,5	64,8(1,88)	35,2	54,8(1,96)	45,2	39,1(1,91)	45,1(1,95)	15,8	30,7(2,63)	35,8(2,74)	33,5	64,8(1,88)	35,2	54,8(1,96)	45,2
IGI_{ML}^2	41,5(1,96)	49,8(1,96)	8,70	33,9(2,70)	36,1(2,74)	29,9	71,2(1,78)	28,8	55,0(1,96)	45,0	41,5(1,96)	49,8(1,96)	8,70	33,9(2,70)	36,1(2,74)	29,9	71,2(1,78)	28,8	55,0(1,96)	45,0
$NON3_{ML}^3$	41,5(1,96)	49,8(1,96)	8,70	34,1(2,71)	36,0(2,74)	29,9	71,1(1,78)	28,9	55,0(1,96)	45,0	41,5(1,96)	49,8(1,96)	8,70	34,1(2,71)	36,0(2,74)	29,9	71,1(1,78)	28,9	55,0(1,96)	45,0

ou \mathbf{m} est une expression vectorielle des estimations par case $\hat{\mathbf{m}}^{fghm} \cdot \hat{\mathbf{\beta}}^{EMP}$ est l'EMP de $\hat{\mathbf{\beta}}$ dont la variance $\text{Var}(\hat{\mathbf{\beta}}^{EMP})$ est donnée par l'inverse de (9), et $\partial \mathbf{m} / \partial \mathbf{\beta} = \mathbf{N}^h \times [\text{diag}(\hat{\pi}) - \hat{\pi} \hat{\pi}^T] \mathbf{Z}$ où $\hat{\pi}$ possède

$$(14) \quad \frac{\partial m'}{\partial p} \text{Var}(\hat{\beta}_{\text{EMP}}) \frac{\partial p}{\partial m}$$

3. Une application à un sondage électoral dans l'Etat de l'Ohio

comme element type.

$$\frac{\sum_{k \in (i,j,h,l,m)} \exp(\mathbf{z}_k^{\text{EMP}})}{\exp(\mathbf{z}_{ijhlm}^{\text{EMP}})} = \pi_{ijhlm}(\hat{\mathbf{g}}^{\text{EMP}})$$

$$N_h \times [\text{diag}(\tilde{\pi}) - \tilde{\pi}\tilde{\pi}']Z \text{ où } \tilde{\pi} \text{ possède}$$

ou \mathbf{m} est une expression vectorielle des estimations par case $\hat{\mathbf{m}}_{\text{ijhlm}}$, $\hat{\beta}_{\text{EMP}}$ est l'EMP de β dont la variance $\text{Var}(\hat{\beta}_{\text{EMP}})$ est donnée par l'inverse de (9), et $\partial \mathbf{m} / \partial \beta =$

Lorsque l'on veut prédire par son sondage le gagnant d'une élection, l'exactitude du résultat dépend souvent de la façon dont sont traités les électeurs indécis qui voteront vraisemblablement, mais qui n'ont pas encore décidé quel est leur candidat préféré. Nous comparons les estimations bayésiennes basées sur les cinq types de lois a priori à l'estimation du MTV en nous servant des données du Buckeye State Poll (BSP), un sondage électoral dans l'Etat de l'Ohio, réalisé en 1998 par le Center for Survey Research de l'Ohio State University. Les sondages préélectoraux du BSP ont produit des tableaux de contingence à double entrée incomplets dont une catégorie était le candidat préféré et l'autre, la probabilité de voter aux élections de novembre 1998 en vue d'élire le gouverneur, le procureur général, le maire de Columbus et le trésorier de l'Ohio. Le tableau 2, qui résume ces quatre sondages, révèle un nombre important d'électeurs indécis.

Pour la comparaison, nous considérons le modèle 1 de non-réponse ignorable et les modèles 2 et 3 de non-réponse non ignorable qui suivent.

Pour la comparaison, nous considérons le modèle 1 de non-réponse ignorable et les modèles 2 et 3 de non-réponse

non ignorable qui suivent.

$$\text{Modèle I : } \log(m_{ijkl}) = \beta_0 + \beta_i^j + \beta_l^k + \beta_{ij}^{kl} + \beta_{ik}^{jl} + \beta_{il}^{jk} +$$

Données observées pour les sondages préélectorales du BSP

Course à l'élection du gouverneur				Course à l'élection du procureur général			
Fisher	Taft	Autre	Indécis	Montgomery	Corday	Indécis	
112	140	23	61	197	82	57	
Ne votera probablement pas	96	108	73	161	65	75	
Indécis	7	11	1	15	4	0	
Course à l'élection du maire				Course à l'élection du trésorier			
Coleman	Teaser	Espy	Indécis	Deters	Donofrio	Indécis	
40	32	25	30	127	119	90	
Votera probablement pas	37	47	56	127	90	84	
Indécis	0	2	1	10	7	0	

.TW IGI

Nous désignons les estimations du MV sous le modèle 1 de non-réponse ignorables, le modèle 2 de non-réponse non ignorable et le modèle 3 de non-réponse non ignorable par IGI_{MV} , $NON2_{NV}$ et $NON3_{NV}$, respectivement. Soit aussi $NON2_{BE}$ l'estimateur bayésien utilisant le type de lois a priori sous le modèle 2. Autrement dit, $NON2_{BE}^1$ utilise les lois a priori axées sur les répondants de (10) et $NON2_{BE}^2$ utilise les mêmes lois a priori que $NON2_{BE}^1$, excepté que $\delta^{(1)} = 0$. De même, $NON2_{BE}^3$ est donné par (11) et $NON2_{BE}^4$ utilise les mêmes lois a priori qu'excepté que $\delta^{(1)} = 0$. $NON2_{BE}^5$ est l'estimation bayésienne obtenue en utilisant les lois a priori constantes de (12). De plus, nous pouvons utiliser la méthode de Stasny (1986, 1988) pour estimer les effectifs prévus par case sous les modèles 1 et 3 qu'elle a supposés implicitement. Cependant, ces estimations semblent être exactement les mêmes que

Dans le modèle 1, les données manquant complètement au hasard et les cas pour lesquels des données manquent peuvent être ignorés dans les inférences de la vraie semblance. Les modèles 2 et 3 sont des modèles de non-réponse non ignorable où la probabilité qu'une variable manquante dépend de la variable elle-même dans le modèle 2, tandis qu'elle dépend de l'autre variable dans le modèle 3. Notons que, sous les modèles 1 et 3, les estimations du MV ne se situent pas sur la limite de l'espace des paramètres comme l'ont montré Baker et coll. (1992). En outre, ayant constaté que, sous les modèles 1 et 3, les cinq estimations bayésiennes des effets attendus par cas sont non seulement assez proches de l'estimation du MV, mais ont aussi presque le même écart-type, nous ne présentons les estimations du MV que pour les modèles 1 et 3.

$$\text{Modèle 3 : } \log(m_{fkl}) = \beta_0 + \beta'_1 X^1_{fj} + \beta'_2 X^2_{fj} + \beta'_3 X^1_{kl} + \beta'_4 X^2_{kl} + \beta'_5 X^1_{R^1_{fj}} + \beta'_6 X^2_{R^1_{fj}} + \beta'_7 X^1_{R^2_{fj}} + \beta'_8 X^2_{R^2_{fj}}.$$

$$\text{Modèle 2 : } \log(m_{fkl}) = \beta_0 + \beta'_1 X^1_l + \beta'_2 X^2_l + \beta'_3 X^3_l + \beta'_4 X^4_l + \beta'_5 X^5_l + \beta'_6 X^6_l + \beta'_7 X^7_l + \beta'_8 X^8_l + \beta'_9 X^9_l + \beta'_{10} X^{10}_l + \beta'_{11} X^{11}_l + \beta'_{12} X^{12}_l + \beta'_{13} X^{13}_l + \beta'_{14} X^{14}_l + \beta'_{15} X^{15}_l$$

Toutefois, il est possible de les étendre afin d'introduire ce genre de pondération. La simple extension qui suit montre comment tenir compte d'une stratification type. Dans un tableau à triple entrée, soit X_3 la troisième variable de réponse marquée de l'indice h ($h = 1, \dots, H$) que nous supposons être toujours observée. Les H catégories peuvent être des strates dans un échantillonnage stratifié. Puisque X_3 est toujours observée, la variable indicatrice de données manquantes correspondante R_3 est égale à 1 et son observation peut être désignée par y^{jhlmi} . Alors, pour chaque strate h , nous pouvons écrire la log-vraisemblance suivante :

$$l_h = \sum_{j=1}^J \sum_{l=1}^L \sum_{i=1}^I \sum_{m=1}^M \log(\pi^{jhl11}) + \sum_{j=1}^J \sum_{l=1}^L \sum_{i=1}^I \sum_{m=1}^M \log(\pi^{jhl12}) + \sum_{j=1}^J \sum_{l=1}^L \sum_{i=1}^I \sum_{m=1}^M \log(\pi^{jhl21}) + \sum_{j=1}^J \sum_{l=1}^L \sum_{i=1}^I \sum_{m=1}^M \log(\pi^{jhl22})$$

où $\pi^{jhlmi} = P(X_1 = i, X_2 = j, R_1 = l, R_2 = m | X_3 = h)$. Donc, la terminologie X_3 utilisée pour un tableau à triple entrée joue le rôle d'un indicateur pour les strates. Pour chaque strate h , la vraisemblance de (13) est exactement la même que celle d'un tableau à double entrée.

Alors, nous pouvons définir un modèle log-linéaire pour l'espérance par case $m^{jhlmi} = N_h \cdot \pi^{jhlmi}$ de la même façon qu'en (2), où $N_h = \sum_{i,j,l,m} y^{jhlmi}$ pour chaque $h = 1, 2, \dots, H$. Un modèle de non-réponse non ignorable est donné par

$$\log(m^{jhlmi}) = \beta_{0h} + \beta_{X_1 h}^{j_1} + \beta_{X_2 h}^{j_2} + \beta_{R_1 h}^{l_1} + \beta_{R_2 h}^{l_2} + \beta_{X_1 X_2 h}^{j_1 j_2} + \beta_{X_1 R_1 h}^{j_1 l_1} + \beta_{X_1 R_2 h}^{j_1 l_2} + \beta_{X_2 R_1 h}^{j_2 l_1} + \beta_{X_2 R_2 h}^{j_2 l_2} + \beta_{X_1 X_2 R_1 h}^{j_1 j_2 l_1} + \beta_{X_1 X_2 R_2 h}^{j_1 j_2 l_2} + \beta_{X_1 R_1 R_2 h}^{j_1 l_1 l_2} + \beta_{X_2 R_1 R_2 h}^{j_2 l_1 l_2} + \beta_{X_1 X_2 R_1 R_2 h}^{j_1 j_2 l_1 l_2}$$

Afin d'éviter le problème des solutions limites décrit à la section 2, nous utilisons les lois a priori de Dirichlet pour π^{jhlmi}

$$\prod_{j=1}^J \prod_{l=1}^L \prod_{i=1}^I \prod_{m=1}^M \pi_{jhlmi}^{\delta_{jhlmi}} \cdot \pi_{jhl21}^{\delta_{jhl21}} \cdot \pi_{jhl22}^{\delta_{jhl22}}$$

Ensuite, nous suivons exactement les mêmes procédures que celles illustrées à la section 2 pour estimer l'espérance par case m^{jhlmi} pour chaque $h = 1, 2, \dots, H$. L'estimation de l'espérance dans la $(i, j)^{\text{e}}$ case est

$$E(y_{ij}) = \sum_{h=1}^H w_h \sum_{l=1}^L \sum_{m=1}^M m^{jhlmi}$$

où w_h est le poids connu pour la h^{e} strate et m^{jhlmi} est le poids pour un échantillon stratifié où N_h est la taille de la population de la h^{e} strate. La matrice de variance-covariance d'une approximation de la distribution de \hat{m} est

Jusqu'ici, nous avons présenté des méthodes pour un tableau à double entrée, et y^{jhlmi} est défini pour l'effectif de la case (i, j) de la j^{e} colonne ('est-à-dire $X_1 = i, X_2 = j$), et l'indicateur R_1 pour une ligne manquante et R_2 pour une colonne manquante ('est-à-dire $R_1 = k, R_2 = l$). L'extension à un tableau à triple entrée est facile. Soit y^{jklm} l'effectif de la $(i, j, k)^{\text{e}}$ case pour les trois variables de réponse ('est-à-dire $X_1 = i, X_2 = j$ et $X_3 = k$) et les lignes et colonnes manquantes respectives ('est-à-dire $R_1 = l, R_2 = m$ et $R_3 = n$ pour $l, m, n = 1, 2$). Donc, $lmn = 111$ implique que chacune des trois variables est observée, $lmn = 112$ implique que X_1 et X_2 sont observées, mais que X_3 manque; de même pour $lmn = 121, 122, 211, 212, 221, 222$, 1 indique que la variable est observée et 2, qu'elle manque. Par conséquent, nous pouvons définir l'algorithme EM et les lois a priori pour un tableau de contingence à triple entrée incomplet. À l'étape E, l'espérance conditionnelle pour la $(i, j, k)^{\text{e}}$ case avec l'information inconnue de la marge k est donnée par

$$E^{\text{old}}(y^{jklm} | \pi^{\text{old}}_{jklm}, y^{j+112}) = y^{j+112} \frac{m^{\text{old}}_{jkl12}}{m^{\text{old}}_{j+112}}$$

De même,

$$E^{\text{old}}(y^{jklm} | \pi^{\text{old}}_{jklm}, y^{j++122}) = y^{j++122} \frac{m^{\text{old}}_{jkl22}}{m^{\text{old}}_{j++122}}$$

et

$$E^{\text{old}}(y^{jklm} | \pi^{\text{old}}_{jklm}, y^{++++22}) = y^{++++22} \frac{m^{\text{old}}_{jkl22}}{m^{\text{old}}_{++++22}}$$

Nous pouvons définir de la même manière d'autres espérances et cinq types de lois a priori.

Le sondage électoral de l'État de l'Ohio est réalisé par la méthode de composition aléatoire (CA). Si le sondage par CA est strictement autopondéré, aucune modification ne doit être apportée aux méthodes bayésiennes (Lavrakas 1993; Pothoff 1994). Cependant, la CA n'est pas toujours exécutée selon un plan autopondéré. Par exemple, un échantillon téléphonique est constitué de ménages et non de personnes. Si une personne est interviewée dans un ménage, la réponse doit être également nécessaire pour les ménages possédant plus d'un numéro de téléphone. Si l'on dispose d'une estimation exacte du nombre total de ménages, on peut procéder à une stratification par région ou par État et il convient d'envisager une pondération dans une analyse complète. Les sondages électoraux réalisés en Ohio en 1998 ont été effectués par CA. Dans la présente étude, nous méthode et modèles ne comprennent pas de pondération pour tenir compte de la stratification, de la mise en grappes et d'autres facteurs dominant lieu à des probabilités de sélection différentes dans un sondage téléphonique.

où b_j est la solution de $\sum_{j=1}^2 y^{j11} b_j = y^{j+12}$, $\hat{m}^{j21} = 0$,

$$\hat{m}^{2/21} = \hat{m}^{2/11} \frac{y^{2+11}}{y^{2+21}}, \hat{m}^{j22} = 0,$$

et $\hat{m}^{j22} = \hat{m}^{2/12} y^{+22} / y^{2+12}$. Par conséquent, ces estimations du MV tiennent compte à la fois de l'information des répondants et des non-répondants. Nous pouvons aussi obtenir les estimations du MV au moyen de notre algorithme EM décrit à la section 2.1 en posant que $\delta_{jkl} = 0$ pour tout i, j, k et l . En utilisant ces estimations du MV, nous définissons le troisième type de loi a priori sous la forme

$$\delta_{j11} = \Delta_{11} \cdot \left(\hat{m}^{j11} \right)^{\delta_{j12}}, \delta_{j12}$$

$$= \Delta_{12} \cdot \left(\hat{m}^{j12} + \frac{I \cdot J}{1} \right) \cdot \frac{1}{2},$$

$$\delta_{j21} = \Delta_{21} \cdot \left(\hat{m}^{j21} + \frac{I \cdot J}{1} \right) \cdot \frac{1}{2},$$

$$\delta_{j22} = \Delta_{22} \cdot \left(\hat{m}^{j22} + \frac{I \cdot J}{1} \right) \cdot \frac{1}{2}$$

et

(11)

$$\delta_{j22} = \frac{3}{J} \cdot \left(\frac{I \cdot J}{1} \right).$$

et

(12)

$$\delta_{j11} = 0, \delta_{j12} = \frac{3}{J} \cdot \left(\frac{I \cdot J}{1} \right), \delta_{j21} = \frac{3}{J} \cdot \left(\frac{I \cdot J}{1} \right),$$

Nous définissons le quatrième type de loi a priori en posant que $\delta_{j11} = 0$ dans (11) comme nous l'avons fait pour obtenir le deuxième type de loi a priori à partir du premier type. Le dernier type de loi a priori, tiré de Clogg et coll. (1991) est défini comme il suit

où $\Delta_{kl} = p \cdot \hat{m}^{k+kl} / \hat{m}^{k+11}$ pour $k, l = 1, 2$, et le terme $1/II$ est la loi a priori constante de Clogg et coll. (1991) afin d'éviter d'éventuelles solutions limites pour m^{j12} , m^{j21} et m^{j22} (voir aussi la cinquième loi a priori de δ_{jkl} que nous attribuons la troisième loi a priori plus loin). Donc, l'EMP de m^{jkl} diminue et se rapproche de l'EMV obtenu sous le modèle de non-réponse non ignorable, tandis que la première loi a priori est obtenue sous le modèle de non-réponse ignorable.

Tableau 1
Cinq types de lois a priori δ_{jkl} (\hat{m}^{jkl} est l'EMV, I et J sont les nombres de lignes et de colonnes dans un tableau à double entrée, et p est le nombre de paramètres)

	δ_{j11}	δ_{j12}	δ_{j21}	δ_{j22}
Type I	$\Delta_{11} \frac{y^{j11}}{y^{j+11}}$	$\Delta_{12} \frac{y^{j11}}{y^{j+11}}$	$\Delta_{21} \frac{y^{j11}}{y^{j+11}}$	$\Delta_{22} \frac{y^{j11}}{y^{j+11}}$
Type II	0	$\Delta_{12} \frac{y^{j11}}{y^{j+11}}$	$\Delta_{21} \frac{y^{j+11}}{y^{j11}}$	$\Delta_{22} \frac{y^{j+11}}{y^{j+11}}$
Type III	$\Delta_{11} \cdot \left(\frac{\hat{m}^{j11}}{\hat{m}^{j+11}} \right)$	$\Delta_{12} \left(\frac{\hat{m}^{j12}}{\hat{m}^{j+12}} + \frac{1}{IJ} \right)$	$\Delta_{21} \left(\frac{\hat{m}^{j21}}{\hat{m}^{j+21}} + \frac{1}{IJ} \right)$	$\Delta_{22} \left(\frac{\hat{m}^{j22}}{\hat{m}^{j+22}} + \frac{1}{IJ} \right)$
Type IV	0	$\Delta_{12} \left(\frac{\hat{m}^{j12}}{\hat{m}^{j+12}} + \frac{1}{IJ} \right)$	$\Delta_{21} \left(\frac{\hat{m}^{j21}}{\hat{m}^{j+21}} + \frac{1}{IJ} \right)$	$\Delta_{22} \left(\frac{\hat{m}^{j22}}{\hat{m}^{j+22}} + \frac{1}{IJ} \right)$
Type V	0	$\Delta_{12} \left(\frac{I \cdot J}{1} \right)$	$\Delta_{21} \left(\frac{I \cdot J}{1} \right)$	$\Delta_{22} \left(\frac{I \cdot J}{1} \right)$
$y^{j+11} = y^{j+11} - y^{j+11}$ et $\hat{m}^{j+11} = \hat{m}^{j+11} - \hat{m}^{j+11}$				

Ces cinq types de lois a priori sont résumés au tableau 1 et sont comparés à la section suivante en utilisant des données empiriques et des études par simulation.

et

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_U - B^{12} & 0 & 0 \\ 0 & 0 & I_U - B^{21} & 0 \\ 0 & 0 & 0 & I_U - B^{22} \end{pmatrix}.$$

Ici, l'axe d'espace et puisque l'extension pour i et j B^{21} et B^{22} généraux n'est pas difficile, nous illustrons B^{12} , B^{21} et B^{22} uniquement pour $I = 2$ et $J = 3$:

$$B^{12} = \begin{pmatrix} m_{112} & m_{112} & m_{112} & 0 & 0 & 0 \\ m_{112} & m_{112} & m_{112} & 0 & 0 & 0 \\ m_{112} & m_{112} & m_{112} & 0 & 0 & 0 \\ m_{112} & m_{112} & m_{112} & 0 & 0 & 0 \\ m_{112} & m_{112} & m_{112} & 0 & 0 & 0 \\ m_{112} & m_{112} & m_{112} & 0 & 0 & 0 \end{pmatrix}$$

$$B^{21} = \begin{pmatrix} m_{121} & 0 & 0 & 0 & 0 & 0 \\ m_{121} & 0 & 0 & 0 & 0 & 0 \\ m_{121} & 0 & 0 & 0 & 0 & 0 \\ m_{121} & 0 & 0 & 0 & 0 & 0 \\ m_{121} & 0 & 0 & 0 & 0 & 0 \\ m_{121} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$B^{22} = \begin{pmatrix} m_{122} & m_{122} & m_{122} & m_{122} & m_{122} & m_{122} \\ m_{122} & m_{122} & m_{122} & m_{122} & m_{122} & m_{122} \\ m_{122} & m_{122} & m_{122} & m_{122} & m_{122} & m_{122} \\ m_{122} & m_{122} & m_{122} & m_{122} & m_{122} & m_{122} \\ m_{122} & m_{122} & m_{122} & m_{122} & m_{122} & m_{122} \\ m_{122} & m_{122} & m_{122} & m_{122} & m_{122} & m_{122} \end{pmatrix}$$

et

Nous constatons que l'information provenant des données observées $\partial^2/\partial\beta^2(Y^{obs})/\partial\beta^2$ est égale à la différence entre l'information provenant des données augmentées et celle provenant des données manquantes. Comme le montrent

2.2 Spécification des lois a priori

L'EMP de β est la variance de l'EMP de β .

Celman, Carlin, Stern et Rubin (2004, page 103), l'inverse de l'information provenant des données observées évalué à l'EMV de β est la variance de l'EMP de β .

Afin d'achever l'algorithme EM, nous devons déterminer les hyperparamètres δ_{jkl} . Nous posons que la somme des lois a priori $\sum_{i,j,k} \delta_{jkl}$ est égale au nombre de paramètres intervenant dans le modèle log-linéaire, p , comme l'ont suggéré Clogit et coll. (1991). Sous cette contrainte, nous proposons cinq types de lois a priori de la façon suivante. Nous commençons par attribuer δ_{jkl} de manière que l'EMP de m_{jkl} diminue et se rapproche de l'EMV obtenu sous non-réponse ignorable. Autrement dit, nous déterminons les lois a priori δ_{jkl} uniquement en fonction des effectifs de réponses connues Y_{jkl}^{111} et nous les appelons lois a priori axées sur les répondants. Le premier type de loi a priori axée sur les répondants est,

$$\delta_{j11} = \Delta_{11}^{Y_{j11}} \delta_{j12} = \Delta_{12}^{Y_{j11}} \delta_{j21} = \Delta_{21}^{Y_{j11}} \delta_{j22} = \Delta_{22}^{Y_{j11}} \quad \text{et} \quad \delta_{j22} = \Delta_{22}^{Y_{j11}} \quad (10)$$

où $\Delta^{kl} = p \cdot Y^{++kl}/Y^{++++}$ pour $k = 1, 2$ et $l = 1, 2$. Le deuxième type de loi a priori axée sur les répondants ne donne aucune loi a priori (c'est-à-dire, comme il est décrit plus bas, qu'aucune loi a priori n'est nécessaire) sur π_{j11} dans le premier type de lois a priori. En d'autres termes, le deuxième type est identique au premier type excepté que $\delta_{j11} = 0$ pour tout j . Dans le cas d'un tableau de contingence à simple entrée (c'est-à-dire que X_1 ou X_2 est entièrement observée sans information manquante) et $Y^{++22} = 0$, le premier type se réduit aux lois a priori utilisées dans Park (1998), tandis que le deuxième type se réduit aux lois a priori utilisées dans Park et Brown (1994). Ces deux types de lois a priori axées sur les répondants pourraient être trop simplistes, parce que l'on suppose généralement que le profil de réponse des non-répondants diffère de celui des répondants sous un modèle de non-réponse non ignorable. Par exemple, le candidat préféré des non-répondants pourrait ne pas être le même que celui des répondants dans un sondage préélectoral.

Afin de définir le troisième type de loi a priori, désignons par m_{jkl}^{EMV} l'EMV de m_{jkl} . Nous pouvons tirer la forme explicite de m_{jkl}^{EMV} de Baker et coll. (1992) où certains m_{jkl}^{EMV} pourraient être nuls à cause des solutions limites. Par exemple, quand une marge supplémentaire de colonne possède une solution limite dans un tableau 2×2 incomplet,

$$m_{211} = Y^{111} = \frac{Y^{2+11}(Y^{211} + Y^{2+21})}{Y^{2+11} + Y^{2+21}}, \quad m_{212} = m_{211} b_j$$

Alors, l'espérance de la fonction a posteriori du log (7)

$$E[\log l_{\text{apos}}] = \sum_{j=1}^I \sum_{i=1}^I y_{ji1}^* \cdot \log(\pi_{ji1})$$

devient

$$+ \sum_{j=1}^I \sum_{i=1}^I y_{ji2}^* \cdot \log(\pi_{ji2})$$

$$+ \sum_{j=1}^I \sum_{i=1}^I y_{ji21}^* \cdot \log(\pi_{ji21})$$

$$+ \sum_{j=1}^I \sum_{i=1}^I y_{ji22}^* \cdot \log(\pi_{ji22}).$$

Cette équation a la même forme que la vraisemblance obtenue à partir d'un tableau de contingence à quatre entrées quand les effectifs par case y_{jil}^* sont entièrement observés. Donc, en utilisant la méthode itérative des moindres carrés repondérés (Agresti 2002, page 342), nous obtenons l'estimateur du maximum a posteriori (EMAP) de β comme il suit :

$$\beta^{(t+1)} = (Z^T Y_{-1}^{-1} Z)^{-1} Y_{-1}^{-1} y^{(t)},$$

où $y^{(t)}$ possède l'élément $y_{ji1}^{(t)} = \log m_{ji1}^{(t)} + (y_{jil}^{(t)} - m_{ji1}^{(t)}) / m_{ji1}^{(t)}$ et $Y_{-1} = [\text{diag}(\mathbf{m}^{(t)})]^{-1}$. Enfin, nous itérons ces étapes E et M jusqu'à ce qu'un critère de convergence soit atteint. Nous choisissons comme critère de convergence $\varepsilon \leq 10^{-6}$, où ε est la différence entre deux fonctions a posteriori du log consécutives.

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \text{diag} \left(\frac{y_{i+12} + \delta_{i+12}}{m_{ji12}} \right) & 0 \\ 0 & \text{diag} \left(\frac{y_{i+21} + \delta_{i+21}}{m_{ji21}} \right) & 0 & 0 \\ 0 & 0 & 0 & \text{diag} \left(\frac{y_{i+22} + \delta_{i+22}}{m_{ji22}} \right) \end{pmatrix}$$

$$\frac{\partial^2 l(\beta | Y_{\text{obs}}, Y_{\text{manq}})}{\partial \beta \partial \beta^T} = \frac{\partial^2 l(\beta | Y_{\text{obs}}, Y_{\text{manq}})}{\partial^2 \log f(Y_{\text{manq}} | Y_{\text{obs}}, \beta)} -$$

Par double dérivation par rapport à β , (8) donne

$$-\log f(Y_{\text{manq}} | Y_{\text{obs}}, \beta) \quad (8)$$

$$l_{\text{pos}} = l(\beta | Y_{\text{obs}}) = l(\beta | Y_{\text{obs}}, Y_{\text{manq}})$$

(5) peut s'écrire

Soit $Y_{\text{obs}} = (y_{ji1}, y_{ji12}, y_{ji21}, y_{ji22})$ pour $i = 1, \dots, I$ et $j = 1, \dots, J$ le vecteur des effectifs observés et le vecteur des effectifs manquants, respectivement. Alors, la log-distribution a posteriori

où π est l'expression vectorielle des probabilités par case π_{jil} et A, B sont données par

$$+ Z^T [\text{diag}(\pi) - \pi \pi^T] A B Z, \quad (9)$$

$$= -Z^T [\text{diag}(\mathbf{m}) - \mathbf{m} \mathbf{m}^T / N] Z$$

$$E^{\text{old}}[l^{\text{apos}}] = \sum_{j=1}^I \sum_{k=1}^I (Y_{j11} + \delta_{j11}) \cdot \log(\pi_{j11}) \\ + \sum_{j=1}^I \sum_{k=1}^I (E^{\text{old}}[Y_{j12} | \pi_{j12}^{\text{old}}, Y_{j+12}] + \delta_{j12}) \cdot \log(\pi_{j12}) \\ + \sum_{j=1}^I \sum_{k=1}^I (E^{\text{old}}[Y_{j21} | \pi_{j21}^{\text{old}}, Y_{j+21}] + \delta_{j21}) \cdot \log(\pi_{j21}) \\ + \sum_{j=1}^I \sum_{k=1}^I (E^{\text{old}}[Y_{j22} | \pi_{j22}^{\text{old}}, Y_{j+22}] + \delta_{j22}) \cdot \log(\pi_{j22}). \quad (7)$$

Puisque Y_{j12} , Y_{j21} et Y_{j22} sont des variables aléatoires multinomiales conditionnées sur les sommes marginales respectives Y_{j+12} , Y_{j+21} et Y_{j+22} , dans l'équation (7), les espérances conditionnelles sont données par

$$E^{\text{old}}(Y_{j12} | \pi_{j12}^{\text{old}}, Y_{j+12}) = Y_{j+12} \frac{m_{j12}^{\text{old}}}{m_{j22}^{\text{old}}}, \\ E^{\text{old}}(Y_{j21} | \pi_{j21}^{\text{old}}, Y_{j+21}) = Y_{j+21} \frac{m_{j21}^{\text{old}}}{m_{j22}^{\text{old}}},$$

et

$$E^{\text{old}}(Y_{j22} | \pi_{j22}^{\text{old}}, Y_{j+22}) = Y_{j+22} \frac{m_{j22}^{\text{old}}}{m_{j22}^{\text{old}}}$$

où $m_{jkl}^{\text{old}} = N \cdot \pi_{jkl}^{\text{old}}$.

Étape M : À cette étape, nous maximisons l'espérance de la loi a posteriori du log (7) en utilisant les pseudo-observations $\tilde{Y}_{j11} = Y_{j11} + \delta_{j11}$, $\tilde{Y}_{j12} = Y_{j12} + \delta_{j12}$, $\tilde{Y}_{j21} = Y_{j21} + \delta_{j21}$ et $\tilde{Y}_{j22} = Y_{j22} + \delta_{j22}$. Nous imposons à ces pseudo-observations que leurs sommes marginales soient égales aux sommes marginales correspondantes des observations $\tilde{Y}_{j+11} = Y_{j+11}$, $\tilde{Y}_{j+12} = Y_{j+12}$, $\tilde{Y}_{j+21} = Y_{j+21}$ et $\tilde{Y}_{j+22} = Y_{j+22}$. Sous ces contraintes, les pseudo-observations sont alors

$$Y_{jkl}^* = \begin{cases} \tilde{Y}_{j11} \frac{Y_{j+11} + \delta_{j+11}}{Y_{j+11}} & \text{pour } k = 1 \text{ et } l = 1 \\ \tilde{Y}_{j12} \frac{Y_{j+12} + \delta_{j+12}}{Y_{j+12}} & \text{pour } k = 1 \text{ et } l = 2 \\ \tilde{Y}_{j21} \frac{Y_{j+21} + \delta_{j+21}}{Y_{j+21}} & \text{pour } k = 2 \text{ et } l = 1 \\ \tilde{Y}_{j22} \frac{Y_{j+22} + \delta_{j+22}}{Y_{j+22}} & \text{pour } k = 2 \text{ et } l = 2. \end{cases}$$

L'équation (5) est assez complexe et nous utilisons donc l'algorithme EM pour estimer les paramètres (c'est-à-dire β).

2.1 L'algorithme EM

Nous maximisons la loi a posteriori donnée en (5) sur le paramètre β en utilisant l'algorithme Espérance-Maximisation généralisé (EMG) (Dempster, Laird et Rubin 1977) comportant les étapes E et M qui suivent.

Étape E : En utilisant les variables augmentées Y_{j12}^* , Y_{j21}^* et Y_{j22}^* pour $i = 1, \dots, I$ et $j = 1, \dots, J$, la loi a posteriori (5) peut s'écrire sous la forme

$$l^{\text{apos}} = \sum_{j=1}^I \sum_{k=1}^I (Y_{j11} + \delta_{j11}) \log(\pi_{j11}) \\ + \sum_{j=1}^I \sum_{k=1}^I (Y_{j12} + \delta_{j12}) \log(\pi_{j12}) \\ + \sum_{j=1}^I \sum_{k=1}^I (Y_{j21} + \delta_{j21}) \log(\pi_{j21}) \\ + \sum_{j=1}^I \sum_{k=1}^I (Y_{j22} + \delta_{j22}) \log(\pi_{j22}). \quad (6)$$

Pour déterminer l'espérance de la loi a posteriori du log donné par (6), nous calculons le moyenne sur les effectifs manquants Y_{j12} , Y_{j21} et Y_{j22} , sachant les estimations courantes des paramètres, π_{jkl}^{old} et les sommes marginales Y_{j+12} , Y_{j+21} et Y_{j+22} :

section, nous calculons également la probabilité de couverture, afin d'examiner la performance des estimations bayésiennes. À la section 5, nous présentons certaines conclusions.

2. Modèles bayésiens

Nous discutons de cinq estimations bayésiennes pour traiter la non-réponse non ignorable dans un tableau de contingence à double entrée incomplet. À la section 2.1, nous présentons un algorithme EM pour aborder le problème de la non-réponse dans un tableau de contingence à double entrée. Puis, à la section 2.2, nous spécifions cinq lois a priori et étendons notre approche à un tableau de contingence à multiples entrées.

Soit X_1 et X_2 ayant pour indice I et J catégories respectivement, dans un tableau de contingence à double entrée. Posons aussi que $R_1 = 1$ quand X_1 est observée et $R_1 = 2$ quand X_1 manque. De même, $R_2 = 1$ quand X_2 est observée et $R_2 = 2$ quand X_2 manque. Alors, la série complète des X_1 , X_2 , R_1 et R_2 produit un tableau de contingence $I \times J \times 2 \times 2$ contenant des effectifs entièrement classifiés, des effectifs partiellement classifiés et des effectifs non classifiés. Afin de distinguer ces trois catégories d'observations, représentons par y^{jkl} l'effectif appartenant à la i^e catégorie de X_1 , la j^e catégorie de X_2 , la k^e valeur de R_1 et la l^e valeur de R_2 . Donc, nous utilisons y^{jkl} pour les marges supplémentaires de colonne et de y^{+j21} et y^{+j22} pour les marges supplémentaires de ligne respectives, et y^{++22} pour les effectifs non classifiés. Nous supposons que ces trois catégories d'observations suivent une loi multinomiale afin d'obtenir la log-vraisemblance suivante :

$$l = \sum_{j=1}^I \sum_{k=1}^J \sum_{l=1}^2 \log(\pi_{jkl}^{y^{jkl}}) + \sum_{j=1}^I \sum_{k=1}^J \log(\pi_{j+21}^{y^{+j21}}) + \sum_{j=1}^I \sum_{k=1}^J \log(\pi_{j+22}^{y^{+j22}}) \quad (1)$$

où $\pi_{jkl}^{y^{jkl}} = \Pr[X_1 = i, X_2 = j, R_1 = k, R_2 = l] = N$ et $\sum_{i,j,k,l} \pi_{jkl}^{y^{jkl}}$ est fixé.

Comme cette fonction de vraisemblance contient plus de paramètres que le nombre de degrés de liberté disponibles pour l'estimation, nous relierons $\pi_{jkl}^{y^{jkl}}$ à des covariables pertinentes en utilisant une fonction log-linéaire. Puisque nous ne disposons pas de variables explicatives, nous n'utilisons aucune. Cependant, des variables explicatives peuvent être intégrées facilement dans le modèle log-linéaire de la même façon qu'y sont intégrées les variables catégoriques (voir Baker et Laird 1988, ainsi que Park et Brown 1994 pour plus de précisions).

Nous définissons un modèle de non-réponse non ignorable pour toutes les variables X_1 , X_2 , R_1 et R_2 par

$$l_{\text{pos}} = \sum_{j=1}^I \sum_{k=1}^J \log(\pi_{jkl}^{y^{jkl}}) - \sum_{j=1}^I \sum_{k=1}^J \sum_{l=1}^2 \log(\pi_{jkl}^{y^{jkl}}) + \sum_{j=1}^I \sum_{k=1}^J \log(\pi_{jkl}^{y^{jkl}}) \quad (2)$$

où les hyperparamètres $\delta_{jkl}^{y^{jkl}}$ sont spécifiés à la section 2.2. Ces lois a priori de Ditcher produisent une forme explicite et comme d'une loi a posteriori, parce qu'ils sont conjugués à une loi multinomiale (Clogg et coll. 1991 ; Park et Brown 1994 ; Forster et Smith 1998). Avec (3), la loi multinomiale de (1) pour les observations et la loi a priori (4), nous obtenons la log-distribution a posteriori :

$$\prod_{j=1}^I \prod_{k=1}^J \prod_{l=1}^2 \pi_{jkl}^{y^{jkl}} \cdot \pi_{j+21}^{y^{+j21}} \cdot \pi_{j+22}^{y^{+j22}} \quad (4)$$

par (2) peut être réécrit sous la forme

$$\log m = Z\beta \quad (3)$$

(2) peut être réécrit sous la forme

(3) peut être réécrit sous la forme

(4) des lois a priori de Ditcher donnés

Afin d'éviter une solution limite de l'estimation du MV dans le modèle (2), nous imposons aux probabilités par case

où le vecteur m de dimensions $I \times J \times 2 \times 2$ est l'espace-
 rance par case et β est la représentation vectorielle des β .
 Afin d'éviter une solution limite de l'estimation du MV dans le modèle (2), nous imposons aux probabilités par case

où le vecteur m de dimensions $I \times J \times 2 \times 2$ est l'espace-
 rance par case et β est la représentation vectorielle des β .
 Afin d'éviter une solution limite de l'estimation du MV dans le modèle (2), nous imposons aux probabilités par case

Ce modèle log-linéaire est saturé, puisque le nombre de paramètres est exactement le même que le nombre de cases observées provenant du tableau de convergence à double entrée incomplet. Il s'agit aussi d'un modèle de non-réponse non ignorable à cause des termes d'interaction entre X_1 et X_2 et entre X_2 et R_2 , ce qui sous-entend que, pour chaque variable de réponse, la non-réponse dépend de la situation de cette variable. Le modèle log-linéaire est un outil utilisé fréquemment pour analyser les tableaux de contingence incomplets avec non-réponse non ignorable. Soit p le nombre de paramètres (c'est-à-dire β) à estimer. Nous introduisons le $p \times 1$ vecteur de plan d'expérience z^{ijkl} pour préciser l'affiliation de l'observation appartenant à la $(i, j, k, l)^e$ catégorie. Alors, le modèle log-linéaire donné en

$$\log(m_{ijkl}) = \beta_0 + \beta_1' X_1 + \beta_2' X_2 + \beta_3' R_1 + \beta_4' R_2 + \beta_5' X_1 X_2 + \beta_6' X_1 R_1 + \beta_7' X_1 R_2 + \beta_8' X_2 R_2 + \beta_9' X_1 X_2 R_1 + \beta_{10}' X_1 X_2 R_2 + \beta_{11}' X_1 R_1 R_2 + \beta_{12}' X_1 R_2 R_2 + \beta_{13}' X_2 R_1 R_2 + \beta_{14}' X_1 X_2 R_1 R_2 + \beta_{15}' X_1 X_2 R_1 R_2 + \beta_{16}' X_1 X_2 R_1 R_2 + \beta_{17}' X_1 X_2 R_1 R_2 + \beta_{18}' X_1 X_2 R_1 R_2 + \beta_{19}' X_1 X_2 R_1 R_2 + \beta_{20}' X_1 X_2 R_1 R_2 + \beta_{21}' X_1 X_2 R_1 R_2 + \beta_{22}' X_1 X_2 R_1 R_2 + \beta_{23}' X_1 X_2 R_1 R_2 + \beta_{24}' X_1 X_2 R_1 R_2 + \beta_{25}' X_1 X_2 R_1 R_2 + \beta_{26}' X_1 X_2 R_1 R_2 + \beta_{27}' X_1 X_2 R_1 R_2 + \beta_{28}' X_1 X_2 R_1 R_2 + \beta_{29}' X_1 X_2 R_1 R_2 + \beta_{30}' X_1 X_2 R_1 R_2 + \beta_{31}' X_1 X_2 R_1 R_2 + \beta_{32}' X_1 X_2 R_1 R_2 + \beta_{33}' X_1 X_2 R_1 R_2 + \beta_{34}' X_1 X_2 R_1 R_2 + \beta_{35}' X_1 X_2 R_1 R_2 + \beta_{36}' X_1 X_2 R_1 R_2 + \beta_{37}' X_1 X_2 R_1 R_2 + \beta_{38}' X_1 X_2 R_1 R_2 + \beta_{39}' X_1 X_2 R_1 R_2 + \beta_{40}' X_1 X_2 R_1 R_2 + \beta_{41}' X_1 X_2 R_1 R_2 + \beta_{42}' X_1 X_2 R_1 R_2 + \beta_{43}' X_1 X_2 R_1 R_2 + \beta_{44}' X_1 X_2 R_1 R_2 + \beta_{45}' X_1 X_2 R_1 R_2 + \beta_{46}' X_1 X_2 R_1 R_2 + \beta_{47}' X_1 X_2 R_1 R_2 + \beta_{48}' X_1 X_2 R_1 R_2 + \beta_{49}' X_1 X_2 R_1 R_2 + \beta_{50}' X_1 X_2 R_1 R_2 + \beta_{51}' X_1 X_2 R_1 R_2 + \beta_{52}' X_1 X_2 R_1 R_2 + \beta_{53}' X_1 X_2 R_1 R_2 + \beta_{54}' X_1 X_2 R_1 R_2 + \beta_{55}' X_1 X_2 R_1 R_2 + \beta_{56}' X_1 X_2 R_1 R_2 + \beta_{57}' X_1 X_2 R_1 R_2 + \beta_{58}' X_1 X_2 R_1 R_2 + \beta_{59}' X_1 X_2 R_1 R_2 + \beta_{60}' X_1 X_2 R_1 R_2 + \beta_{61}' X_1 X_2 R_1 R_2 + \beta_{62}' X_1 X_2 R_1 R_2 + \beta_{63}' X_1 X_2 R_1 R_2 + \beta_{64}' X_1 X_2 R_1 R_2 + \beta_{65}' X_1 X_2 R_1 R_2 + \beta_{66}' X_1 X_2 R_1 R_2 + \beta_{67}' X_1 X_2 R_1 R_2 + \beta_{68}' X_1 X_2 R_1 R_2 + \beta_{69}' X_1 X_2 R_1 R_2 + \beta_{70}' X_1 X_2 R_1 R_2 + \beta_{71}' X_1 X_2 R_1 R_2 + \beta_{72}' X_1 X_2 R_1 R_2 + \beta_{73}' X_1 X_2 R_1 R_2 + \beta_{74}' X_1 X_2 R_1 R_2 + \beta_{75}' X_1 X_2 R_1 R_2 + \beta_{76}' X_1 X_2 R_1 R_2 + \beta_{77}' X_1 X_2 R_1 R_2 + \beta_{78}' X_1 X_2 R_1 R_2 + \beta_{79}' X_1 X_2 R_1 R_2 + \beta_{80}' X_1 X_2 R_1 R_2 + \beta_{81}' X_1 X_2 R_1 R_2 + \beta_{82}' X_1 X_2 R_1 R_2 + \beta_{83}' X_1 X_2 R_1 R_2 + \beta_{84}' X_1 X_2 R_1 R_2 + \beta_{85}' X_1 X_2 R_1 R_2 + \beta_{86}' X_1 X_2 R_1 R_2 + \beta_{87}' X_1 X_2 R_1 R_2 + \beta_{88}' X_1 X_2 R_1 R_2 + \beta_{89}' X_1 X_2 R_1 R_2 + \beta_{90}' X_1 X_2 R_1 R_2 + \beta_{91}' X_1 X_2 R_1 R_2 + \beta_{92}' X_1 X_2 R_1 R_2 + \beta_{93}' X_1 X_2 R_1 R_2 + \beta_{94}' X_1 X_2 R_1 R_2 + \beta_{95}' X_1 X_2 R_1 R_2 + \beta_{96}' X_1 X_2 R_1 R_2 + \beta_{97}' X_1 X_2 R_1 R_2 + \beta_{98}' X_1 X_2 R_1 R_2 + \beta_{99}' X_1 X_2 R_1 R_2 + \beta_{100}' X_1 X_2 R_1 R_2 + \beta_{101}' X_1 X_2 R_1 R_2 + \beta_{102}' X_1 X_2 R_1 R_2 + \beta_{103}' X_1 X_2 R_1 R_2 + \beta_{104}' X_1 X_2 R_1 R_2 + \beta_{105}' X_1 X_2 R_1 R_2 + \beta_{106}' X_1 X_2 R_1 R_2 + \beta_{107}' X_1 X_2 R_1 R_2 + \beta_{108}' X_1 X_2 R_1 R_2 + \beta_{109}' X_1 X_2 R_1 R_2 + \beta_{110}' X_1 X_2 R_1 R_2 + \beta_{111}' X_1 X_2 R_1 R_2 + \beta_{112}' X_1 X_2 R_1 R_2 + \beta_{113}' X_1 X_2 R_1 R_2 + \beta_{114}' X_1 X_2 R_1 R_2 + \beta_{115}' X_1 X_2 R_1 R_2 + \beta_{116}' X_1 X_2 R_1 R_2 + \beta_{117}' X_1 X_2 R_1 R_2 + \beta_{118}' X_1 X_2 R_1 R_2 + \beta_{119}' X_1 X_2 R_1 R_2 + \beta_{120}' X_1 X_2 R_1 R_2 + \beta_{121}' X_1 X_2 R_1 R_2 + \beta_{122}' X_1 X_2 R_1 R_2 + \beta_{123}' X_1 X_2 R_1 R_2 + \beta_{124}' X_1 X_2 R_1 R_2 + \beta_{125}' X_1 X_2 R_1 R_2 + \beta_{126}' X_1 X_2 R_1 R_2 + \beta_{127}' X_1 X_2 R_1 R_2 + \beta_{128}' X_1 X_2 R_1 R_2 + \beta_{129}' X_1 X_2 R_1 R_2 + \beta_{130}' X_1 X_2 R_1 R_2 + \beta_{131}' X_1 X_2 R_1 R_2 + \beta_{132}' X_1 X_2 R_1 R_2 + \beta_{133}' X_1 X_2 R_1 R_2 + \beta_{134}' X_1 X_2 R_1 R_2 + \beta_{135}' X_1 X_2 R_1 R_2 + \beta_{136}' X_1 X_2 R_1 R_2 + \beta_{137}' X_1 X_2 R_1 R_2 + \beta_{138}' X_1 X_2 R_1 R_2 + \beta_{139}' X_1 X_2 R_1 R_2 + \beta_{140}' X_1 X_2 R_1 R_2 + \beta_{141}' X_1 X_2 R_1 R_2 + \beta_{142}' X_1 X_2 R_1 R_2 + \beta_{143}' X_1 X_2 R_1 R_2 + \beta_{144}' X_1 X_2 R_1 R_2 + \beta_{145}' X_1 X_2 R_1 R_2 + \beta_{146}' X_1 X_2 R_1 R_2 + \beta_{147}' X_1 X_2 R_1 R_2 + \beta_{148}' X_1 X_2 R_1 R_2 + \beta_{149}' X_1 X_2 R_1 R_2 + \beta_{150}' X_1 X_2 R_1 R_2 + \beta_{151}' X_1 X_2 R_1 R_2 + \beta_{152}' X_1 X_2 R_1 R_2 + \beta_{153}' X_1 X_2 R_1 R_2 + \beta_{154}' X_1 X_2 R_1 R_2 + \beta_{155}' X_1 X_2 R_1 R_2 + \beta_{156}' X_1 X_2 R_1 R_2 + \beta_{157}' X_1 X_2 R_1 R_2 + \beta_{158}' X_1 X_2 R_1 R_2 + \beta_{159}' X_1 X_2 R_1 R_2 + \beta_{160}' X_1 X_2 R_1 R_2 + \beta_{161}' X_1 X_2 R_1 R_2 + \beta_{162}' X_1 X_2 R_1 R_2 + \beta_{163}' X_1 X_2 R_1 R_2 + \beta_{164}' X_1 X_2 R_1 R_2 + \beta_{165}' X_1 X_2 R_1 R_2 + \beta_{166}' X_1 X_2 R_1 R_2 + \beta_{167}' X_1 X_2 R_1 R_2 + \beta_{168}' X_1 X_2 R_1 R_2 + \beta_{169}' X_1 X_2 R_1 R_2 + \beta_{170}' X_1 X_2 R_1 R_2 + \beta_{171}' X_1 X_2 R_1 R_2 + \beta_{172}' X_1 X_2 R_1 R_2 + \beta_{173}' X_1 X_2 R_1 R_2 + \beta_{174}' X_1 X_2 R_1 R_2 + \beta_{175}' X_1 X_2 R_1 R_2 + \beta_{176}' X_1 X_2 R_1 R_2 + \beta_{177}' X_1 X_2 R_1 R_2 + \beta_{178}' X_1 X_2 R_1 R_2 + \beta_{179}' X_1 X_2 R_1 R_2 + \beta_{180}' X_1 X_2 R_1 R_2 + \beta_{181}' X_1 X_2 R_1 R_2 + \beta_{182}' X_1 X_2 R_1 R_2 + \beta_{183}' X_1 X_2 R_1 R_2 + \beta_{184}' X_1 X_2 R_1 R_2 + \beta_{185}' X_1 X_2 R_1 R_2 + \beta_{186}' X_1 X_2 R_1 R_2 + \beta_{187}' X_1 X_2 R_1 R_2 + \beta_{188}' X_1 X_2 R_1 R_2 + \beta_{189}' X_1 X_2 R_1 R_2 + \beta_{190}' X_1 X_2 R_1 R_2 + \beta_{191}' X_1 X_2 R_1 R_2 + \beta_{192}' X_1 X_2 R_1 R_2 + \beta_{193}' X_1 X_2 R_1 R_2 + \beta_{194}' X_1 X_2 R_1 R_2 + \beta_{195}' X_1 X_2 R_1 R_2 + \beta_{196}' X_1 X_2 R_1 R_2 + \beta_{197}' X_1 X_2 R_1 R_2 + \beta_{198}' X_1 X_2 R_1 R_2 + \beta_{199}' X_1 X_2 R_1 R_2 + \beta_{200}' X_1 X_2 R_1 R_2 + \beta_{201}' X_1 X_2 R_1 R_2 + \beta_{202}' X_1 X_2 R_1 R_2 + \beta_{203}' X_1 X_2 R_1 R_2 + \beta_{204}' X_1 X_2 R_1 R_2 + \beta_{205}' X_1 X_2 R_1 R_2 + \beta_{206}' X_1 X_2 R_1 R_2 + \beta_{207}' X_1 X_2 R_1 R_2 + \beta_{208}' X_1 X_2 R_1 R_2 + \beta_{209}' X_1 X_2 R_1 R_2 + \beta_{210}' X_1 X_2 R_1 R_2 + \beta_{211}' X_1 X_2 R_1 R_2 + \beta_{212}' X_1 X_2 R_1 R_2 + \beta_{213}' X_1 X_2 R_1 R_2 + \beta_{214}' X_1 X_2 R_1 R_2 + \beta_{215}' X_1 X_2 R_1 R_2 + \beta_{216}' X_1 X_2 R_1 R_2 + \beta_{217}' X_1 X_2 R_1 R_2 + \beta_{218}' X_1 X_2 R_1 R_2 + \beta_{219}' X_1 X_2 R_1 R_2 + \beta_{220}' X_1 X_2 R_1 R_2 + \beta_{221}' X_1 X_2 R_1 R_2 + \beta_{222}' X_1 X_2 R_1 R_2 + \beta_{223}' X_1 X_2 R_1 R_2 + \beta_{224}' X_1 X_2 R_1 R_2 + \beta_{225}' X_1 X_2 R_1 R_2 + \beta_{226}' X_1 X_2 R_1 R_2 + \beta_{227}' X_1 X_2 R_1 R_2 + \beta_{228}' X_1 X_2 R_1 R_2 + \beta_{229}' X_1 X_2 R_1 R_2 + \beta_{230}' X_1 X_2 R_1 R_2 + \beta_{231}' X_1 X_2 R_1 R_2 + \beta_{232}' X_1 X_2 R_1 R_2 + \beta_{233}' X_1 X_2 R_1 R_2 + \beta_{234}' X_1 X_2 R_1 R_2 + \beta_{235}' X_1 X_2 R_1 R_2 + \beta_{236}' X_1 X_2 R_1 R_2 + \beta_{237}' X_1 X_2 R_1 R_2 + \beta_{238}' X_1 X_2 R_1 R_2 + \beta_{239}' X_1 X_2 R_1 R_2 + \beta_{240}' X_1 X_2 R_1 R_2 + \beta_{241}' X_1 X_2 R_1 R_2 + \beta_{242}' X_1 X_2 R_1 R_2 + \beta_{243}' X_1 X_2 R_1 R_2 + \beta_{244}' X_1 X_2 R_1 R_2 + \beta_{245}' X_1 X_2 R_1 R_2 + \beta_{246}' X_1 X_2 R_1 R_2 + \beta_{247}' X_1 X_2 R_1 R_2 + \beta_{248}' X_1 X_2 R_1 R_2 + \beta_{249}' X_1 X_2 R_1 R_2 + \beta_{250}' X_1 X_2 R_1 R_2 + \beta_{251}' X_1 X_2 R_1 R_2 + \beta_{252}' X_1 X_2 R_1 R_2 + \beta_{253}' X_1 X_2 R_1 R_2 + \beta_{254}' X_1 X_2 R_1 R_2 + \beta_{255}' X_1 X_2 R_1 R_2 + \beta_{256}' X_1 X_2 R_1 R_2 + \beta_{257}' X_1 X_2 R_1 R_2 + \beta_{258}' X_1 X_2 R_1 R_2 + \beta_{259}' X_1 X_2 R_1 R_2 + \beta_{260}' X_1 X_2 R_1 R_2 + \beta_{261}' X_1 X_2 R_1 R_2 + \beta_{262}' X_1 X_2 R_1 R_2 + \beta_{263}' X_1 X_2 R_1 R_2 + \beta_{264}' X_1 X_2 R_1 R_2 + \beta_{265}' X_1 X_2 R_1 R_2 + \beta_{266}' X_1 X_2 R_1 R_2 + \beta_{267}' X_1 X_2 R_1 R_2 + \beta_{268}' X_1 X_2 R_1 R_2 + \beta_{269}' X_1 X_2 R_1 R_2 + \beta_{270}' X_1 X_2 R_1 R_2 + \beta_{271}' X_1 X_2 R_1 R_2 + \beta_{272}' X_1 X_2 R_1 R_2 + \beta_{273}' X_1 X_2 R_1 R_2 + \beta_{274}' X_1 X_2 R_1 R_2 + \beta_{275}' X_1 X_2 R_1 R_2 + \beta_{276}' X_1 X_2 R_1 R_2 + \beta_{277}' X_1 X_2 R_1 R_2 + \beta_{278}' X_1 X_2 R_1 R_2 + \beta_{279}' X_1 X_2 R_1 R_2 + \beta_{280}' X_1 X_2 R_1 R_2 + \beta_{281}' X_1 X_2 R_1 R_2 + \beta_{282}' X_1 X_2 R_1 R_2 + \beta_{283}' X_1 X_2 R_1 R_2 + \beta_{284}' X_1 X_2 R_1 R_2 + \beta_{285}' X_1 X_2 R_1 R_2 + \beta_{286}' X_1 X_2 R_1 R_2 + \beta_{287}' X_1 X_2 R_1 R_2 + \beta_{288}' X_1 X_2 R_1 R_2 + \beta_{289}' X_1 X_2 R_1 R_2 + \beta_{290}' X_1 X_2 R_1 R_2 + \beta_{291}' X_1 X_2 R_1 R_2 + \beta_{292}' X_1 X_2 R_1 R_2 + \beta_{293}' X_1 X_2 R_1 R_2 + \beta_{294}' X_1 X_2 R_1 R_2 + \beta_{295}' X_1 X_2 R_1 R_2 + \beta_{296}' X_1 X_2 R_1 R_2 + \beta_{297}' X_1 X_2 R_1 R_2 + \beta_{298}' X_1 X_2 R_1 R_2 + \beta_{299}' X_1 X_2 R_1 R_2 + \beta_{300}' X_1 X_2 R_1 R_2 + \beta_{301}' X_1 X_2 R_1 R_2 + \beta_{302}' X_1 X_2 R_1 R_2 + \beta_{303}' X_1 X_2 R_1 R_2 + \beta_{304}' X_1 X_2 R_1 R_2 + \beta_{305}' X_1 X_2 R_1 R_2 + \beta_{306}' X_1 X_2 R_1 R_2 + \beta_{307}' X_1 X_2 R_1 R_2 + \beta_{308}' X_1 X_2 R_1 R_2 + \beta_{309}' X_1 X_2 R_1 R_2 + \beta_{310}' X_1 X_2 R_1 R_2 + \beta_{311}' X_1 X_2 R_1 R_2 + \beta_{312}' X_1 X_2 R_1 R_2 + \beta_{313}' X_1 X_2 R_1 R_2 + \beta_{314}' X_1 X_2 R_1 R_2 + \beta_{315}' X_1 X_2 R_1 R_2 + \beta_{316}' X_1 X_2 R_1 R_2 + \beta_{317}' X_1 X_2 R_1 R_2 + \beta_{318}' X_1 X_2 R_1 R_2 + \beta_{319}' X_1 X_2 R_1 R_2 + \beta_{320}' X_1 X_2 R_1 R_2 + \beta_{321}' X_1 X_2 R_1 R_2 + \beta_{322}' X_1 X_2 R_1 R_2 + \beta_{323}' X_1 X_2 R_1 R_2 + \beta_{324}' X_1 X_2 R_1 R_2 + \beta_{325}' X_1 X_2 R_1 R_2 + \beta_{326}' X_1 X_2 R_1 R_2 + \beta_{327}' X_1 X_2 R_1 R_2 + \beta_{328}' X_1 X_2 R_1 R_2 + \beta_{329}' X_1 X_2 R_1 R_2 + \beta_{330}' X_1 X_2 R_1 R_2 + \beta_{331}' X_1 X_2 R_1 R_2 + \beta_{332}' X_1 X_2 R_1 R_2 + \beta_{333}' X_1 X_2 R_1 R_2 + \beta_{334}' X_1 X_2 R_1 R_2 + \beta_{335}' X_1 X_2 R_1 R_2 + \beta_{336}' X_1 X_2 R_1 R_2 + \beta_{337}' X_1 X_2 R_1 R_2 + \beta_{338}' X_1 X_2 R_1 R_2 + \beta_{339}' X_1 X_2 R_1 R_2 + \beta_{340}' X_1 X_2 R_1 R_2 + \beta_{341}' X_1 X_2 R_1 R_2 + \beta_{342}' X_1 X_2 R_1 R_2 + \beta_{343}' X_1 X_2 R_1 R_2 + \beta_{344}' X_1 X_2 R_1 R_2 + \beta_{345}' X_1 X_2 R_1 R_2 + \beta_{346}' X_1 X_2 R_1 R_2 + \beta_{347}' X_1 X_2 R_1 R_2 + \beta_{348}' X_1 X_2 R_1 R_2 + \beta_{349}' X_1 X_2 R_1 R_2 + \beta_{350}' X_1 X_2 R_1 R_2 + \beta_{351}' X_1 X_2 R_1 R_2 + \beta_{352}' X_1 X_2 R_1 R_2 + \beta_{353}' X_1 X_2 R_1 R_2 + \beta_{354}' X_1 X_2 R_1 R_2 + \beta_{355}' X_1 X_2 R_1 R_2 + \beta_{356}' X_1 X_2 R_1 R_2 + \beta_{357}' X_1 X_2 R_1 R_2 + \beta_{358}' X_1 X_2 R_1 R_2 + \beta_{359}' X_1 X_2 R_1 R_2 + \beta_{360}' X_1 X_2 R_1 R_2 + \beta_{361}' X_1 X_2 R_1 R_2 + \beta_{362}' X_1 X_2 R_1 R_2 + \beta_{363}' X_1 X_2 R_1 R_2 + \beta_{364}' X_1 X_2 R_1 R_2 + \beta_{365}' X_1 X_2 R_1 R_2 + \beta_{366}' X_1 X_2 R_1 R_2 + \beta_{367}' X_1 X_2 R_1 R_2 + \beta_{368}' X_1 X_2 R_1 R_2 + \beta_{369}' X_1 X_2 R_1 R_2 + \beta_{370}' X_1 X_2 R_1 R_2 + \beta_{371}' X_1 X_2 R_1 R_2 + \beta_{372}' X_1 X_2 R_1 R_2 + \beta_{373}' X_1 X_2 R_1 R_2 + \beta_{374}' X_1 X_2 R_1 R_2 + \beta_{375}' X_1 X_2 R_1 R_2 + \beta_{376}' X_1 X_2 R_1 R_2 + \beta_{377}' X_1 X_2 R_1 R_2 + \beta_{378}' X_1 X_2 R_1 R_2 + \beta_{379}' X_1 X_2 R_1 R_2 + \beta_{380}' X_1 X_2 R_1 R_2 + \beta_{381}' X_1 X_2 R_1 R_2 + \beta_{382}' X_1 X_2 R_1 R_2 + \beta_{383}' X_1 X_2 R_1 R_2 + \beta_{384}' X_1 X_2 R_1 R_2 + \beta_{385}' X_1 X_2 R_1 R_2 + \beta_{386}' X_1 X_2 R_1 R_2 + \beta_{387}' X_1 X_2 R_1 R_2 + \beta_{388}' X_1 X_2 R_1 R_2 + \beta_{389}' X_1 X_2 R_1 R_2 + \beta_{390}' X_1 X_2 R_1 R_2 + \beta_{391}' X_1 X_2 R_1 R_2 + \beta_{392}' X_1 X_2 R_1 R_2 + \beta_{393}' X_1 X_2 R_1 R_2 + \beta_{394}' X_1 X_2 R_1 R_2 + \beta_{395}' X_1 X_2 R_1 R_2 + \beta_{396}' X_1 X_2 R_1 R_2 + \beta_{397}' X_1 X_2 R_1 R_2 + \beta_{398}' X_1 X_2 R_1 R_2 + \beta_{399}' X_1 X_2 R_1 R_2 + \beta_{400}' X_1 X_2 R_1 R_2 + \beta_{401}' X_1 X_2 R_1 R_2 + \beta_{402}' X_1 X_2 R_1 R_2 + \beta_{403}' X_1 X_2 R_1 R_2 + \beta_{404}' X_1 X_2 R_1 R_2 + \beta_{405}' X_1 X_2 R_1 R_2 + \beta_{406}' X_1 X_2 R_1 R_2 + \beta_{407}' X_1 X_2 R_1 R_2 + \beta_{408}' X_1 X_2 R_1 R_2 + \beta_{409}' X_1 X_2 R_1 R_2 + \beta_{410}' X_1 X_2 R_1 R_2 + \beta_{411}' X_1 X_2 R_1 R_2 + \beta_{412}' X_1 X_2 R_1 R_2 + \beta_{413}' X_1 X_2 R_1 R_2 + \beta_{414}' X_1 X_2 R_1 R_2 + \beta_{415}' X_1 X_2 R_1 R_2 + \beta_{416}' X_1 X_2 R_1 R_2 + \beta_{417}' X_1 X_2 R_1 R_2 + \beta_{418}' X_1 X_2 R_1 R_2 + \beta_{419}' X_1 X_2 R_1 R_2 + \beta_{420}' X_1 X_2 R_1 R_2 + \beta_{421}' X_1 X_2 R_1 R_2 + \beta_{422}' X_1 X_2 R_1 R_2 + \beta_{423}' X_1 X_2 R_1 R_2 + \beta_{424}' X_1 X_2 R_1 R_2 + \beta_{425}' X_1 X_2 R_1 R_2 + \beta_{426}' X_1 X_2 R_1 R_2 + \beta_{427}' X_1 X_2 R_1 R_2 + \beta_{428}' X_1 X_2 R_1 R_2 + \beta_{429}' X_1 X_2 R_1 R_2 + \beta_{430}' X_1 X_2 R_1 R_2 + \beta_{431}' X_1 X_2 R_1 R_2 + \beta_{432}' X_1 X_2 R_1 R_2 + \beta_{433}' X_1 X_2 R_1 R_2 + \beta_{434}' X_1 X_2 R_1 R_2 + \beta_{435}' X_1 X_2 R_1 R_2 + \beta_{436}' X_1 X_2 R_1 R_2 + \beta_{437}' X_1 X_2 R_1 R_2 + \beta_{438}' X_1 X_2 R_1 R_2 + \beta_{439}' X_1 X_2 R_1 R_2 + \beta_{440}' X_1 X_2 R_1 R_2 + \beta_{441}' X_1 X_2 R_1 R_2 + \beta_{442}' X_1 X_2 R_1 R_2 + \beta_{443}' X_1 X_2 R_1 R_2 + \beta_{444}' X_1 X_2 R_1 R_2 + \beta_{445}' X_1 X_2 R_1 R_2 + \beta_{446}' X_1 X_2 R_1 R_2 + \beta_{447}' X_1 X_2 R_1 R_2 + \beta_{448}' X_1 X_2 R_1 R_2 + \beta_{449}' X_1 X_2 R_1 R_2 + \beta_{450}' X_1 X_2 R_1 R_2 + \beta_{451}' X_1 X_2 R_1 R_2 + \beta_{452}' X_1 X_2 R_1 R_2 + \beta_{453}' X_1 X_2 R_1 R_2 + \beta_{454}' X_1 X_2 R_1 R_2 + \beta_{455}' X_1 X_2 R_1 R_2 + \beta_{456}' X_1 X_2 R_1 R_2 + \beta_{457}' X_1 X_2 R_1 R_2 + \beta_{458}' X_1 X_2 R_1 R_2 + \beta_{459}' X_1 X_2 R_1 R_2 + \beta_{460}' X_1 X_2 R_1 R_2 + \beta_{461}' X_1 X_2 R_1 R_2 + \beta_{462}' X_1 X_2 R_1 R_2 + \beta_{463}' X_1 X_2 R_1 R_2 + \beta_{464}' X_1 X_2 R_1 R_2 + \beta_{465}' X_1 X_2 R_1 R_2 + \beta_{466}' X_1 X_2 R_1 R_2 + \beta_{467}' X_1 X_2 R_1 R_2 + \beta_{468}' X_1 X_2 R_1 R_2 + \beta_{469}' X_1 X_2 R_1 R_2 + \beta_{470}' X_1 X_2 R_1 R_2 + \beta_{471}' X_1 X_2 R_1 R_2 + \beta_{472}' X_1 X_2 R_1 R_2 + \beta_{473}' X_1 X_2 R_1 R_2 + \beta_{474}' X_1 X_2 R_1 R_2 + \beta_{475}' X_1 X_2 R_1 R_2 + \beta_{476}' X_1 X_2 R_1 R_2 + \beta_{477}' X_1 X_2 R_1 R_2 + \beta_{478}' X_1 X_2 R_1 R_2 + \beta_{479}' X_1 X_2 R_1 R_2 + \beta_{480}' X_1 X_2 R_1 R_2 + \beta_{481}' X_1 X_2 R_1 R_2 + \beta_{482}' X_1 X_2 R_1 R_2 + \beta_{483}' X_1 X_2 R_1 R_2 + \beta_{484}' X_1 X_2 R_1 R_2 + \beta_{485}' X_1 X_2 R_1 R_2 + \beta_{486}' X_1 X_2 R_1 R_2 + \beta_{487}' X_1 X_2 R_1 R_2 + \beta_{488}' X_1 X_2 R_1 R_2 + \beta_{489}' X_1 X_2 R_1 R_2 + \beta_{490}' X_1 X_2 R_1 R_2 + \beta_{491}' X_1 X_2 R_1 R_2 + \beta_{492}' X_1 X_2 R_1 R_2 + \beta_{493}' X_1 X_2 R_1 R_2 + \beta_{494}' X_1 X_2 R_1 R_2 + \beta_{495}' X_1 X_2 R_1 R_2 + \beta_{496}' X_1 X_2 R_1 R_2 + \beta_{497}' X_1 X_2 R_1 R_2 + \beta_{498}' X_1 X_2 R_1 R_2 + \beta_{499}' X_1 X_2 R_1 R_2 + \beta_{500}' X_$$

fonction de vraisemblance. Dans ce cas, les estimations du maximum de vraisemblance (MV) des paramètres du modèle log-linéaire ne peuvent pas avoir de solution unique et présentent habituellement de grands écarts-types (voir la section 4 ou Baker, Rosenberger et Dersimonian (1992), ainsi que Park et Brown (1994) pour des discussions plus détaillées).

Les conditions dans lesquelles l'estimation du MV se situe sur la solution limite ont été proposées dans le cas d'un tableau de contingence à simple entrée (Baker et Laird 1988; Michiels et Moltenbergs 1997). L'explication géométrique de la solution limite de l'estimation du MV a été présentée (Smith et coll. 1999; Clark 2002). Baker et coll. (1992) ont énoncé une condition suffisante et nécessaire sous laquelle l'estimation du MV peut avoir une solution limite dans un tableau de contingence à double entrée.

Afin de contourner ce genre de problème de solution limite dans l'estimation du MV en présence de non-réponse non ignorable, Park et Brown (1994) et Park (1998) ont proposé une approche bayésienne au moyen de lois a priori empiriques basées uniquement sur l'information fournie par les répondants. Clogg, Rubin, Schenker et Schultz (1991) ont utilisé une loi a priori constante pour un tableau de contingence à simple entrée incomplet. Ils ont montré qu'en cas de non-réponse non ignorable, les méthodes bayésiennes donnent des erreurs quadratiques moyennes (EQM) plus petites que l'estimation du MV pour les espérances par cas, mais notre étude par simulation révèle que cela n'est généralement pas vérifié dans un tableau de contingence à double entrée incomplet. Donc, nous présentons deux modèles bayésiens dont les lois a priori dépendent de l'information provenant à la fois des répondants et des électeurs indécis. Puis, nous appliquons chacun d'eux pour analyser le tableau de contingence à double entrée incomplet. L'extension à un tableau à multiples entrées est simple. Nous pouvons appliquer facilement cette extension à des données pondérées issues d'un échantillonnage stratifié ou en grappes en utilisant les covariables appropriées (voir la section 2.2).

La suite de l'article est divisée en quatre sections. À la section 2, nous considérons des modèles bayésiens avec cinq lois a priori différentes et présentons un algorithme d'Espérance-Maximisation (EM) généralisé pour estimer les probabilités par cas. À la section 3, nous appliquons les modèles bayésiens à quatre jeux de données empiriques provenant du sondage électoral de l'Etat de l'Ohio et nous comparons les estimations bayésiennes à l'estimation du MV ainsi qu'aux résultats réels de l'élection. À la section 4, nous recourons à des études par simulation pour comparer les EQM et les biais des estimations bayésiennes fondées sur différents pourcentages de données manquantes et profils de réponse des répondants et des non-répondants. Dans cette

(1994), l'erreur de prédiction des résultats réels de l'élection augmente parallèlement au taux d'électeurs indécis. Afin de surmonter ce problème, Monterola, Lim, Garcia et Saloma (2001) ont adopté une approche basée sur des réseaux neuronaux pour classer les électeurs indécis dans un sondage d'opinion publique. Smith, Skinner et Clarke (1999) et Moltenbergs, Kenward et Goetghebuer (2001) ont utilisé des méthodes d'imputation fondées sur un modèle pour le British General Election Survey de 1992 et pour le sondage d'opinion publique réalisé à l'occasion du plébiscite de 1991 en Slovaquie. Notre objectif principal étant d'obtenir des prédictions plus exactes grâce à l'atténuation des électeurs indécis aux cas appropriés, nous avons traité ces derniers comme des observations manquantes à l'instar des chercheurs susmentionnés.

Les non-réponses (ou, de manière équivalente, les électeurs indécis) peuvent être réparties en trois catégories (Little et Rubin 2002, page 11), à savoir les données manquantes complètement au hasard (MCAR pour *missing completely at random*), ce qui signifie que la probabilité d'une non-réponse pour une variable d'intérêt est indépendante de toute variable étudiée, y compris la variable elle-même, les données manquantes au hasard (MAR pour *missing at random*), ce qui signifie que la probabilité d'une non-réponse dépend uniquement des données observées, et les données ne manquant pas au hasard (MNAR pour *missing not at random*), ce qui signifie que la probabilité d'une non-réponse dépend des valeurs non observées. Les modèles qui correspondent aux cas MCAR ou MAR sont appelés modèles de non-réponse ignorable, tandis que ceux correspondant aux cas MNAR sont appelés modèles de non-réponse non ignorable. Par exemple, dans un sondage préélectoral, si les personnes qui répondent n'indiquent pas leur préférence pour un candidat bien qu'elles appuient un candidat particulier, le profil des préférences pour les candidats pourrait ne pas être le même pour les répondants et les non-répondants. Dans ces conditions, le mécanisme de non-réponse est non ignorable. Si l'on suppose que le mécanisme de création des données manquantes est MCAR, l'effet de la non-réponse peut être éliminé dans l'inférence de la vraisemblance (Little et Rubin 2002, page 11). Cependant, si le profil de réponse des non-répondants diffère de celui des répondants, le fait d'écartier les non-réponses ou de spécifier incorrectement le mécanisme de création de la non-réponse donne lieu à des estimations entachées d'une plus grande variance et d'un plus grand biais (Chen 1972; Park et Brown 1994).

Dans les tableaux de contingence, si la non-réponse est non ignorable, l'estimation du MV donne fréquemment des solutions limites où il est estimé que la probabilité de non-réponse est nulle dans certaines cases. Ces solutions limites fournissent souvent un maximum local de la

Méthodes bayésiennes pour un tableau de contingence à double entrée incomplet avec application aux sondages électoraux de l'État de l'Ohio

Bo-Seung Choi, Jai Won Choi et Yousung Park¹

Résumé

Nous appliquons une méthode bayésienne pour résoudre le problème des solutions limitées de l'estimation du maximum de vraisemblance (MV) dans un tableau de contingence à double entrée incomplet en utilisant un modèle log-linéaire et des lois a priori de Dirichlet. Nous comparons cinq lois a priori de Dirichlet pour estimer les probabilités multinomiales par cas sous un modèle de non-réponse non ignorable. Trois de ces lois a priori ont été utilisées dans le cas d'un tableau à simple entrée incomplet et les deux autres sont deux nouvelles lois a priori proposées afin de tenir compte de la différence entre les profils de réponse des répondants et des électeurs indécis. Les estimations bayésiennes obtenues à l'aide des trois premières lois a priori n'ont pas systématiquement de meilleures propriétés que les estimations du MV, contrairement à ce qu'indiquaient des études antérieures, tandis que les deux nouvelles lois a priori donnent de meilleurs résultats que les trois lois a priori antérieures et que les estimations du MV chaque fois qu'est obtenue une solution limite. Nous utilisons quatre jeux de données provenant des sondages électoraux réalisés en 1998 dans l'État de l'Ohio pour illustrer comment il convient d'utiliser et d'interpréter les résultats des estimations pour les élections. Nous procédons à des études par simulation pour comparer les propriétés de cinq estimations bayésiennes sous un modèle de non-réponse non ignorable.

Mots clés : Analyse bayésienne ; non-réponse non ignorable ; tableau de contingence ; solution limite ; algorithme EM.

1. Introduction

Fréquente dans la plupart des sondages, la non-réponse devient un problème sérieux à mesure que son taux augmente (De Heer 1999 ; Groves et Couper 1998). Si l'on résume des données de sondage dans un tableau de contingence à double entrée, ce dernier contient des effectifs de case entièrement classifiés, des effectifs partiellement classifiés (c'est-à-dire non-réponses partielles) et des effectifs non classifiés (c'est-à-dire non-réponses totales). Par exemple, dans le sondage électoral de l'Ohio (Chen et Siasny 2003), l'une des catégories est la préférence de l'électeur (candidat A, B ou C, ou indécis) et l'autre est la probabilité de voter (votera vraisemblablement, ne votera vraisemblablement pas, et indécis). La première marge supplémentaire contient uniquement des données sur les préférences des électeurs, la deuxième contient uniquement des données sur la probabilité de voter et la troisième donne uniquement le nombre de non-réponses totales (cas où les deux réponses sont inconnues). Nous désirons intégrer ces observations manquantes dans l'estimation de l'appui réel pour chaque candidat et présenter des modèles bayésiens pour prédire le gagnant.

Dans certains sondages, les réponses indécises sont traitées comme une catégorie de réponse valide si les répondants ne manifestent pas de préférence nette pour un candidat ni d'intention nette de voter (Smith 1984 ; Rubin, Stern et Vehovar 1995). Néanmoins, de nombreuses études

ont révélé que le comportement de vote des électeurs indécis peut avoir une incidence importante sur le résultat final et que l'on peut améliorer l'exactitude de la prédiction des résultats de l'élection en tenant compte de ces électeurs (Perry 1979 ; Fenwick, Wiseman, Becker et Heiman 1982 ; Myers et O'Connor 1983 ; Kim 1995 ; Chen et Siasny 2003 ; Martin, Traugott et Kennedy 2005). Parmi ces auteurs, Perry (1979) a montré au moyen de données empiriques obtenues selon une approche de scrutin secret que le pourcentage d'électeurs indécis observé dans un sondage est vraisemblablement plus élevé que le pourcentage réel. Kim (1995) a également indiqué que le rôle de ces électeurs indécis est critique, surtout quand leur nombre est supérieur à l'écart entre les deux candidats en tête dans une course électorale. Trois de nos études empiriques décrites à la section 3 appartiennent à ce cas critique. Fenwick et coll. (1982) ainsi que Kim (1995) ont procédé à une analyse discriminante des données du sondage électoral d'octobre 1980 au Massachusetts et des données de l'élection présidentielle américaine de 1992 d'après laquelle ils ont réparti les électeurs indécis entre les candidats afin de montrer qu'en général, ces électeurs indécis ne se rendent pas aux urnes dans les mêmes proportions que les candidats aux urnes décidés. Si l'on met l'accent sur le candidat homologues décidés, les réponses indécises comme des données manquantes (Myers et O'Connor 1983). Comme l'ont mentionné Flannelly, Flannelly et McLeod (2000) et Lau

1. Bo-Seung Choi, directeur de recherche, Institut d'économie, Université de la Corée, Séoul 136-701 ; Yousung Park, professeur, Département de statistique, Université de la Corée, Séoul 136-701, Corée. Courriel : yspark@korea.ac.kr.

Kovar, J.G., Rao, J.N.K., et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, Supplément, 25-45.

Langlet, É.R., Faucher, D. et Lesage, É. (2003). An application of the bootstrap variance estimation method to the Canadian Participation and Activity Limitation Survey. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2299-2306.

Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

Rao, J.N.K., et Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: An appraisal and update. *Analysis of survey data*, (Eds. R.L. Chambers et C.J. Skinner). New-York : John Wiley & Sons, Inc.

Seber, G.A.F. (1984). *Multivariate Observations*. New-York : John Wiley & Sons, Inc.

Shao, J., et Tu, D. (1995). *The jackknife and the bootstrap*. New-York : Springer-Verlag.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, 20, 135-154.

Thomas, D.R., et Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Thomas, D.R., Singh, A.C. et Roberts, G.R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64, 295-311.

Yeo, D., Mantel, H. et Liu, T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

double bootstrap) est estimé par la méthode bootstrap comme dans (5.1). Un double bootstrap nécessite-rait la création d'un autre ensemble de répliques bootstrap pour chaque réplique bootstrap initiale. Il permettrait peut-être d'obtenir de meilleures procédures de test, mais risquerait de ne pas être commode pour les analystes. En élan axé sur des statistiques moins compliquées qui ne requièrent pas de procédure bootstrap, notre méthode de test évite le double bootstrap et reste simple.

Les propriétés de notre méthode dépendent non seulement du choix de la statistique de test, mais aussi de la construction des poids bootstrap. Habituellement, ceux-ci reflètent les deux premiers moments de l'erreur d'échantillonnage sous le plan, ce qui devrait suffire dans la plupart des cas à satisfaire nos hypothèses bootstrap 3, 4 et 5. Les poids bootstrap qui reflètent aussi le troisième moment sous le plan seraient peut-être utiles pour améliorer le niveau d'exactitude du test bootstrap. Cette question doit être étudiée plus en profondeur. Enfin, comme nous l'avons déjà mentionné à la section 3, les poids bootstrap standard fondés sur le plan ne satisfont l'hypothèse 5 que si la fraction d'échantillonnage globale est négligeable de sorte que la partie de la variance totale (3.1) due au modèle est négligeable. Les travaux doivent se poursuivre en vue d'élaborer des poids bootstrap appropriés si la fraction d'échantillonnage est non négligeable qui refléteront les composantes dues au modèle et au plan de la variance totale.

Remerciements

Nous remercions sincèrement le rédacteur associé et deux examinateurs de leurs commentaires. Nous remercions aussi J.N.K. Rao, de l'Université Carleton, ainsi que David Binder et Yves LaFortune, de Statistique Canada, de leurs commentaires et discussions intéressantes sur le sujet. Tous ces commentaires et discussions nous ont permis d'améliorer la qualité générale et la clarté de l'article.

Annexe

Preuve du résultat 1

Partant de l'hypothèse 1, nous voyons facilement que

$$\sqrt{n}(\mathbf{H}\hat{\beta}^{ws} - \mathbf{H}\beta) \xrightarrow{mp} N(0, \mathbf{H}\Sigma\mathbf{H}'). \quad (\text{A.1})$$

En utilisant un résultat standard sur des formes quadratiques (par exemple, Seber 1984, page 540) et l'équation (A.1), nous obtenons

$$n(\mathbf{H}\hat{\beta}^{ws} - \mathbf{H}\beta)' \mathbf{A}^{-1}(\mathbf{H}\hat{\beta}^{ws} - \mathbf{H}\beta) \xrightarrow{mp} \sum_{q=1}^q \lambda_q \Omega_q, \quad (\text{A.2})$$

Bibliographie

Benjamin, Y., et Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Séries B*, 57, 289-300.

Binder, D.A., et Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of survey data*, (Éds. R.L. Chambers et C.J. Skinner). New-York : John Wiley & Sons, Inc.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Efron, B., et Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York : Chapman & Hall.

Far, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

Felleger, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

Graubard, B.I., et Korn, E.L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.

Graubard, B.I., Korn, E.L. et Midtune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 170-174.

Hall, P., et Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757-762.

Hughes, A.L., et Brodsky, M.D. (1994). Variance estimation of drug abuse episodes using the bootstrap. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 212-217.

Korn, E.L., et Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *The American Statistician*, 44, 270-276.

Korn, E.L., et Graubard, B.I. (1991). A note on the large sample properties of linearization, jackknife and balanced repeated replication methods for stratified samples. *The Annals of Statistics*, 19, 2275-2279.

Tableau 4B
Taux de rejet au seuil de signification de 5 % sous SCHEME_EQUAL et échantillonnage non informatif

SCHEME_EQUAL											
Échantillonnage non informatif						Échantillonnage non informatif					
Ho VRAIE			Ho FAVUSSE			Ho VRAIE			Ho FAVUSSE		
$\alpha_1 = 0$			$\alpha_1 = 0,25$			$\alpha_1 = 0,50$			$\alpha_1 = 0,75$		
Test1	Test2	Test1	Test1	Test2	Test1	Test1	Test2	Test1	Test1	Test2	Test1
Naïve non pondérée	4,9	4,5	17,2	12,4	54,3	42,2	88,2	81,7	88,3	81,9	81,7
Wald	5,0	4,5	17,4	12,5	54,7	42,7	88,3	81,9	88,9	85,0	82,6
Rao-Scott	5,7	5,0	18,8	13,1	56,6	49,2	88,9	82,6	88,9	81,8	81,8
Bootstrap proposé	5,0	3,3	16,4	10,0	53,2	36,5	86,8	77,6	88,9	81,8	81,8

Le tableau 4B contient les résultats sous le scénario SCHEME_EQUAL avec échantillonnage non informatif. Dans ce tableau, les méthodes ne semblent pas différer spectaculairement. Comme prévu, la méthode naïve (versions pondérée et non pondérée) donne de bons résultats

proposée reste légèrement conservatrice sous ce scénario non informatif et est légèrement moins puissante que les Bootstron dans cette étude par simulation. La méthode proposée ne supasse pas les méthodes de Rao-Scott et de

Pour étudier l'effet de grands échantillons sur les méthodes de test, nous avons également exécuté certaines simulations avec des tailles d'échantillon dix fois plus grandes que dans les conditions originales, comme l'a

proposé un examinateur. Nous avons considéré une taille de population de 100 000 et sélectionné 1 000 échantillons de la même faible fraction d'échantillonnage. Dans ces conditions, nous avons obtenu les résultats quand H_0 est vraie

présentées au tableau 5 pour l'échantillonnage informatif ainsi que non informatif sous répartition inégale entre les strates. Comme nous y attendions, toutes les méthodes

autres que les méthodes naïves ont produit des taux de rejet semblables qui étaient effectivement légèrement inférieurs à 5 %. Ces résultats montrent que les écarts entre les méthodes deviennent moins importants quand la taille

Dans l'ensemble, la méthode bootstrap que nous proposons était la meilleure en ce qui concerne le niveau, suivie de près par la méthode de Rao-Scott. Elle a toutefois donné des résultats un peu plus conservateurs sous les scénarios d'échantillonnage non informatif, ce qui a été accompagné d'une légère perte de puissance. La méthode de Rao-Scott est une bonne alternative si les utilisateurs ont simple à utiliser, mais est parfois trop libérale et la méthode

niveau ou la puissance, quoique la méthode naïve non pondérée soit viable si l'on est raisonnablement certain que l'échantillonnage n'est pas informatif.

Tableau 5
Taux de rejet au seuil de signification de 5 % sous SCHEME_UNEQUAL

SCHEME_UNEQUAL											
SCHEME_UNEQUAL						SCHEME_UNEQUAL					
Informatif			Ho VRAIE			Non informatif			Ho VRAIE		
$\alpha_1 = 0$			$\alpha_1 = 0$			$\alpha_1 = 0$			$\alpha_1 = 0$		
Test1	Test2	Test1	Test1	Test2	Test1	Test1	Test2	Test1	Test1	Test2	Test1
Naïve non pondérée	100,0	100,0	3,7	3,8	10,5	9,3	3,7	3,8	10,5	9,3	3,8
Naïve pondérée	1,3	0,7	4,6	4,5	3,2	4,1	3,2	4,1	3,2	3,2	3,8
Wald	4,6	4,5	4,6	3,8	3,2	3,8	3,2	3,8	3,2	3,2	3,6
Rao-Scott	4,6	4,5	4,6	3,8	3,2	3,8	3,2	3,8	3,2	3,2	3,6
Bootstron	4,6	4,5	4,6	3,8	3,2	3,8	3,2	3,8	3,2	3,2	3,6
Bootstrap proposé	4,4	3,6	2,9	3,8	3,8	3,8	2,9	3,8	3,8	2,9	3,8

7. Sommaire et discussion

Nous avons proposé une méthode bootstrap générale et simple de test d'hypothèses à partir des données d'enquête qui pourrait également être appliquée dans d'autres domaines que celui des sondages. Notre méthode s'appuie sur des statistiques classiques de test fondées sur un modèle, si bien qu'il est facile pour les analystes de l'appliquer en utilisant des logiciels classiques. Nous avons montré au moyen d'une étude par simulation qu'elle donne de bons résultats dans le contexte d'un modèle de régression linéaire. Ces bons résultats sont encourageants et laissent peut-être entendre que la méthode bootstrap que nous proposons pourrait être utile dans le cas d'autres modèles plus complexes et d'autres statistiques. L'idée pourrait également être adaptée facilement à la construction d'intervalles de confiance bootstrap.

Nous pourrions également envisager de traiter par le bootstrap une statistique asymptotiquement pivot, telle que celle de Rao-Scott (5.4). Cela comporterait cependant un

du niveau, le taux de rejet étant de 6,2 % quand H_0 est vraie. Cependant, la méthode de Rao-Scott est un peu plus puissante que la méthode bootstrap proposée.

Le tableau 3B contient les résultats sous le scénario SCHEME_EQUAL dans le cas de l'échantillonnage informatif. Ici, les versions pondérée et non pondérée de la méthode naïve produisent des résultats semblables, puisque la variabilité des poids de sondage est assez faible. Même dans ce cas, la méthode naïve est définitivement trop conservatrice, ce qui se traduit par une puissance extrêmement faible. Toutes les autres méthodes sont comparables en ce qui concerne le niveau (H_0 vraie) et la puissance (H_0 fausse), quoique la méthode de Wald soit encore légèrement trop libérale comparativement à celles de Bonferroni et de Rao-Scott, ainsi qu'à la méthode bootstrap proposée, avec un taux de rejet de 7,9 % quand H_0 est vraie.

Tableau 3A
Taux de rejet au seuil de signification de 5 % sous SCHEME_UNEQUAL et échantillonnage informatif

Méthode	$\alpha_1 = 0$		$\alpha_1 = 0,25$		$\alpha_1 = 0,50$		$\alpha_1 = 0,75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
SCHEME_UNEQUAL	Échantillonnage informatif							
Ho VRAIE	Ho FAUSSE							
Naïve non pondérée	37,5	100,0	85,3	100,0	100,0	100,0	100,0	100,0
Wald	1,7	0,4	14,5	4,6	58,0	33,6	90,3	78,6
Naïve pondérée	8,0	15,8	30,9	37,1	71,8	73,9	93,1	95,4
Rao-Scott	8,0	6,8	30,9	21,1	71,8	61,7	93,1	91,8
Bonferroni	8,0	11,4	30,9	32,6	71,8	68,8	93,1	91,9
Bootstrap proposé	7,4	6,2	29,4	19,7	70,2	59,7	92,8	91,0

Tableau 3B
Taux de rejet au seuil de signification de 5 % sous SCHEME_EQUAL et échantillonnage informatif

Méthode	$\alpha_1 = 0$		$\alpha_1 = 0,25$		$\alpha_1 = 0,50$		$\alpha_1 = 0,75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
SCHEME_EQUAL	Échantillonnage informatif							
Ho VRAIE	Ho FAUSSE							
Naïve non pondérée	0,1	0,0	6,7	0,3	58,1	16,5	97,2	79,7
Wald	0,1	0,0	6,3	0,3	56,8	18,2	97,0	81,4
Naïve pondérée	5,8	7,9	43,6	37,5	93,7	92,3	99,9	100,0
Rao-Scott	5,8	5,5	43,6	32,1	93,7	90,4	99,9	99,9
Bonferroni	5,8	6,2	43,6	33,6	93,7	88,6	99,9	99,8
Bootstrap proposé	2,3	5,1	42,3	31,0	93,6	89,6	99,9	99,9

Tableau 4A
Taux de rejet au seuil de signification de 5 % sous SCHEME_UNEQUAL et échantillonnage non informatif

Méthode	$\alpha_1 = 0$		$\alpha_1 = 0,25$		$\alpha_1 = 0,50$		$\alpha_1 = 0,75$	
	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
SCHEME_UNEQUAL	Échantillonnage non informatif							
Ho VRAIE	Ho FAUSSE							
Naïve non pondérée	4,2	4,7	13,5	11,2	39,9	34,6	71,8	70,5
Wald	7,6	8,6	16,8	17,8	42,9	42,6	72,5	76,2
Rao-Scott	7,6	6,4	16,8	12,3	42,9	32,1	72,5	72,5
Bonferroni	7,6	7,1	16,8	16,5	42,9	42,1	72,5	75,0
Bootstrap proposé	6,3	4,5	14,4	9,2	38,5	26,4	68,2	56,4

SUDAAN, version 9. La statistique de Bonferroni (5.2) est également obtenue au moyen de SUDAAN. La méthode proposée est programmée dans le logiciel statistique SAS, version 8.

De plus, nous avons exécuté l'étude par simulation en utilisant l'estimateur de variance linéarisé dans les méthodes de Wald, de Rao-Scott et de Bonferroni au lieu de l'estimateur de variance bootstrap (5.1). Les taux de rejet obtenus par échantillonnage aléatoire simple stratifié avec $n_h - 1$ tirages à partir des n_h unités d'échantillonnage dans la strate h . Cette méthode tient compte de la variabilité liée au plan d'échantillonnage (avec une légère surestimation de la variance par rapport au plan due à l'hypothèse d'échantillonnage avec remise), mais ignore la variabilité du modèle, ce qui est acceptable, puisque la fraction globale d'échantillonnage (1/100) est faible.

6.5 Résultats de simulation

Pour chaque population, scénario de stratification, scénario de répartition, hypothèse nulle et méthode, nous avons calculé le taux de rejet en pourcentage sur les 5 000 échantillons sélectionnés (en utilisant un seuil de signification de 5 %). Les résultats sont présentés ci-après dans les tableaux 3A, 3B, 4A et 4B. Ils sont plus frappants et plus intéressants pour l'hypothèse nulle TEST2 que pour l'hypothèse TEST1. Par conséquent, nous axerons notre discussion sur les résultats concernant la première.

Les tableaux 3A et 3B contiennent les résultats dans le cas de l'échantillonnage informatif, qui présente un plus grand intérêt pour nous. Commentons par examiner les résultats du tableau 3A pour SCHEME_UNEQUAL. Les deux méthodes naïves ont donné des résultats médiocres car elles n'exploitent pas convenablement l'information du plan d'échantillonnage. D'une part, la version non pondérée est définitivement trop libérale, car son taux de rejet est de loin supérieur à 5 % sous l'hypothèse nulle. D'autre part, la version pondérée est trop conservatrice et manque nettement de puissance comparativement aux autres méthodes. La méthode de Wald est trop libérale, le taux de rejet étant de 15,8 % quand H_0 est vraie. La méthode simple de Bonferroni améliore la situation, bien qu'elle soit encore trop libérale, avec un taux de rejet de 11,4 % quand H_0 est vraie. Ce résultat est un peu surprenant, car la méthode de Bonferroni a la réputation d'être (asymptotiquement) conservatrice. Un examinateur a suggéré que nous envisagions une méthode Bonferroni améliorée, telle que celle élaborée par Benjamini et Hochberg (1995). Dans la présente étude par simulation, ce genre de méthode ne serait pas utile, car elle produit systématiquement un taux de rejet plus élevé que la méthode de Bonferroni standard. La méthode de Rao-Scott l'emporte significativement sur les méthodes de Wald et de Bonferroni sous l'hypothèse nulle avec un taux de rejet de 6,8 %. La méthode bootstrap proposée est comparable à la méthode éprouvée, mais plus compliquée de Rao-Scott, avec peut-être même une légère amélioration

Notons que nous définissons simplement le poids de sondage w_k comme étant l'inverse de la probabilité de sélection de l'unité k .

Enfin, pour chaque échantillon sélectionné, nous avons calculé 500 poids bootstrap fondés sur le plan pour chaque unité échantillonnée, comme l'ont décrit Rao et coll. (1992),

entre autres. Dans notre application de cette méthode, chaque échantillon bootstrap a été sélectionné avec remise par échantillonnage aléatoire simple stratifié avec $n_h - 1$ tirages à partir des n_h unités d'échantillonnage dans la strate h . Cette méthode tient compte de la variabilité liée au plan d'échantillonnage (avec une légère surestimation de la variance par rapport au plan due à l'hypothèse d'échantillonnage avec remise), mais ignore la variabilité du modèle, ce qui est acceptable, puisque la fraction globale d'échantillonnage (1/100) est faible.

6.3 Hypothèses nulles

Pour chaque échantillon sélectionné, nous avons modélisé y_k sous forme d'une fonction de x_k en nous servant d'un modèle d'analyse de variance. Plus précisément, nous avons défini les variables indicatrices

$$x_k = \begin{cases} 1, & \text{si } y_k = I, \\ 0, & \text{autrement,} \end{cases}$$

pour $i = 1, \dots, I$, et ajusté le modèle linéaire $y_k = \beta_0 + \sum_{i=1}^I \beta_i x_k + \varepsilon_k$ en utilisant la méthode des moindres carrés pondérés, où ε_k est un terme d'erreur aléatoire de moyenne nulle et de variance constante. Nous avons envisagé de tester les deux hypothèses nulles suivantes :

$$\begin{aligned} \text{TEST1 : } H_0 : \beta_1 &= 0 \\ \text{TEST2 : } H_0 : \beta_1 &= \beta_2 = \dots = \beta_{I-1} = 0. \end{aligned}$$

Notons que ces hypothèses sont toutes deux vraies pour la population obtenue avec $\alpha_1 = 0$, tandis qu'elles sont fausses pour les autres populations. Les trois dernières populations sont utilisées pour évaluer la puissance des diverses méthodes de test étudiées.

6.4 Méthodes de test

Pour chaque échantillon sélectionné, nous avons testé les deux hypothèses nulles susmentionnées avec les cinq méthodes suivantes : la méthode bootstrap proposée, la méthode naïve (versions non pondérée et pondérée) décrite à la section 5.1, la méthode de Bonferroni décrite à la section 5.2, la méthode F de Wald décrite à la section 5.3 et la méthode F de Rao-Scott décrite à la section 5.4. Les résultats pour la méthode naïve sont la sortie standard du logiciel SAS, tandis que les statistiques F de Wald et de Rao-Scott sont les sorties standard du logiciel statistique

Nous comparons la statistique $t_{\text{RS}}^w(s; c)$ à la queue supérieure de la distribution $F_{\tilde{Q}, d - \tilde{Q} + 1}$. Cette procédure est exécutée au moyen du logiciel SUDAN (Research Triangle Institute 2004).

5.4 Méthode F de Rao-Scott

Une autre méthode consiste à utiliser une version F (voir Rao et Thomas 2003) de la statistique khi-carré ajustée de deuxième ordre de Rao et Scott (1981), qui est fondée sur la correction de Satterthwaite pour le nombre de degrés de liberté. Nous utilisons une adaptation de la méthode de ces auteurs pour la régression linéaire, comme elle est exécutée dans le logiciel SUDAN (Research Triangle Institute 2004). La statistique est définie par

$$t_{\text{RS}}^w(s; c) = \frac{\lambda(1 + a^2) \tilde{O}^*}{1}$$

$$(\mathbf{H}\hat{\beta}_{\text{RS}}^w - c)'(\mathbf{H}'\mathbf{V}_{\text{RS}}^w(\hat{\beta}_{\text{RS}}^w) \mathbf{H})^{-1}(\mathbf{H}\hat{\beta}_{\text{RS}}^w - c), (5.4)$$

où $\mathbf{V}_{\text{RS}}^w(\hat{\beta}_{\text{RS}}^w)$ est un estimateur de la matrice de variance-covariance de $\hat{\beta}_{\text{RS}}^w$ sous un plan d'échantillonnage aléatoire simple, λ est la moyenne des valeurs propres de la matrice d'effets de plans généralisés $[\mathbf{V}_{\text{RS}}^w(\hat{\beta}_{\text{RS}}^w)]^{-1} \mathbf{V}_{\text{RS}}^w(\hat{\beta}_{\text{RS}}^w)$, a est le coefficient de variation de ces valeurs propres et $\tilde{O}^* = \tilde{O}/(1 + a^2)$. La statistique F de Rao-Scott $t_{\text{RS}}^w(s; c)$ est comparée à la queue supérieure de la distribution $F_{\tilde{Q}, d}$.

6. Étude par simulation

Nous avons effectué une étude par simulation pour examiner le niveau et la puissance des méthodes de test susmentionnées dans le cas d'un échantillonnage informatif ainsi que non informatif. Aux sections 6.1 et 6.2, nous décrivons la création des populations et des échantillons, doit être testée à la section 6.3, décrivons les méthodes évaluées à la section 6.4 et présentons les résultats de la simulation à la section 6.5.

6.1 Création des populations

Nous avons créé quatre populations de $N = 10\,000$ unités. Pour commencer, nous avons créé indépendamment une variable catégorique v_k pour chaque unité de population k telle que $v_k = i$, pour $i = 1, \dots, I$, avec la probabilité $P(v_k = i) = 1/I$, où I est le nombre de catégories de v_k , qui a été fixé à 5. La variable dépendante y a été créée selon

$$y_k = \alpha_0 + \alpha_1 \left(v_k - \frac{I+1}{2} \right) + \sigma \varphi_k, (6.1)$$

où $\varphi_k \sim N(0, 1)$, $\alpha_0 = 10$ et $\sigma = 3$. Les quatre populations que nous avons créées ne diffèrent que par le

choix de α_1 , qui contrôle la corrélation entre y et v . Nous avons considéré les valeurs $\alpha_1 = 0, 0,25, 0,50$ et $0,75$.

6.2 Création des échantillons et des poids bootstrap

Pour chacune des quatre populations susmentionnées, nous avons sélectionné 5 000 échantillons aléatoires simples de taille 100 sans remise sous deux scénarios de stratification différents destinés à simuler l'échantillonnage informatif, ainsi que non informatif. Dans le cas de l'échantillonnage non informatif, les strates correspondent exactement aux cinq catégories de la variable v définies plus haut. Dans le cas de l'échantillonnage informatif, les strates sont définies par classification croisée de la variable v et d'une autre variable catégorique z qui dépend du terme d'erreur aléatoire $\sigma \varphi_k$ dans (6.1). Pour chaque unité de population k , nous avons créé la variable z de la façon suivante : $z_k = 1$, si $\sigma \varphi_k > 0$, et $z_k = 2$, autrement. Dans le cas informatif, cela donne dix strates qui sont construites par recoupement entre les cinq catégories de v et les deux catégories de z . Chacune des dix strates informatives contient environ 1 000 unités de population, tandis que chacune des cinq strates non informatives contient environ 2 000 unités de population. En outre, nous avons utilisé deux scénarios distincts de répartition aux strates. Sous le scénario SCHEME_UNEQUAL, les 100 unités échantillonnées sont réparties entre les strates de la façon suivante :

Tableau 1
Tailles d'échantillon pour SCHEME_UNEQUAL

Informatif		Non informatif	
v	z	v	z
1	1	1	1
2	2	3	2
3	3	4	3
4	4	5	4
5	5		5

Sous le second scénario, désigné SCHEME_EQUAL, le même nombre d'unité est affecté à chaque strate de la façon suivante :

Tableau 2
Tailles d'échantillon pour SCHEME_EQUAL

Informatif		Non informatif	
v	z	v	z
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5

Ces scénarios produisent deux ensembles très différents de poids d'échantillonnage. Les poids résultant de la répartition SCHEME_UNEQUAL sont beaucoup plus variables que ceux obtenus sous le scénario SCHEME_EQUAL.

$$\hat{V}^{mp}(\hat{\beta}^{ws}) = \frac{B}{\sum_{b=1}^B (\hat{\beta}^{ws,b} - \hat{\beta}^{ws})(\hat{\beta}^{ws,b} - \hat{\beta}^{ws})'} \quad (5.1)$$

Il convient de souligner que la validité de l'hypothèse 5 est donc requise non seulement pour la méthode bootstrap que nous proposons, mais aussi pour les méthodes de Bonferroni, Wald et Rao-Scott.

5.1 Deux méthodes naïves

La version pondérée de la méthode naïve consiste à utiliser la statistique $\hat{t}(s, \mathbf{w}^s; \mathbf{c})$ donnée par (4.2), qui est comparée à la queue supérieure de la distribution $F_{Q,n-r}^{\hat{t}(s, \mathbf{w}^s; \mathbf{c})}$. La version non pondérée s'appuie sur la statistique $\hat{t}(s, \mathbf{I}_s; \mathbf{c})$, qui est de nouveau comparée à la queue supérieure de la distribution $F_{Q,n-r}^{\hat{t}(s, \mathbf{I}_s; \mathbf{c})}$. Bien qu'on ne puisse pas s'attendre à ce que ces deux méthodes donnent de bons résultats sous échantillonnage informatif, elles sont néanmoins utilisées fréquemment en pratique, surtout la version pondérée. Notons que si l'échantillonnage n'est pas complet du plan version non pondérée, qui ne tient pas compte du plan d'échantillonnage, aboutit à un test simple, valide et raisonnablement puissant.

5.2 Méthode de Bonferroni

La méthode de Bonferroni a été étudiée par Korn et Graubard (1990). Elle est simple à appliquer et ces auteurs ont montré qu'elle donnait de bons résultats dans leur étude empirique. Pour décrire cette procédure, représentons par \mathbf{H}^q la q^e ligne de \mathbf{H} et par c_q^s le q^e élément de \mathbf{c} . Ensuite, calculons les \hat{Q} statistiques pondérées

$$\hat{t}_{BON}^q(s; \mathbf{c}^q) = \frac{\mathbf{H}^q \hat{\mathbf{V}}^{mp}(\hat{\beta}^{ws}) \mathbf{H}^q}{\mathbf{H}^q \hat{\beta}^{ws} - c_q^s} \quad (5.2)$$

Nous comparons la plus grande statistique $\hat{t}_{BON}^q(s; \mathbf{c}^q)$, pour $q = 1, \dots, \hat{Q}$, à la queue supérieure de la distribution $F_{1,d}$ avec un seuil de signification révisé α / \hat{Q} au lieu de α . Le nombre de degrés de liberté d est égal au nombre d'unités primaires d'échantillonnage échantillonnées moins le nombre de strates. Notons que cette méthode dépend en général de la paramétrisation du modèle utilisé.

5.3 Méthode F de Wald

Nous pouvons définir une version F de la statistique khi-carré de Wald standard avec nombre de degrés de liberté ajusté au dénominateur comme l'a proposé Fellegi (1980) sous la forme

$$\hat{F}^{mp}(s; \mathbf{c}) = \frac{d - \hat{Q} + 1}{d} \frac{\hat{Q} \hat{\mathbf{V}}^{mp}(\hat{\beta}^{ws}) \mathbf{H}^q (\mathbf{H}^q \hat{\beta}^{ws} - c_q^s)}{\mathbf{H}^q \hat{\beta}^{ws} - c_q^s} \quad (5.3)$$

Remarque 2 : Définissons la statistique bootstrap $\hat{t}(s, \mathbf{w}^s; \mathbf{c}^k)$ en remplaçant y_k par $e_k = y_k - \mathbf{x}_k^t \hat{\beta}^{ws}$ dans $\hat{t}(s, \mathbf{w}^s; \mathbf{0})$, pour chaque $k \in s$. Il n'est pas difficile de montrer que $\hat{t}(s, \mathbf{w}^s; \mathbf{0}) = \hat{t}(s, \mathbf{w}^s; \mathbf{c}^k)$ de sorte que notre méthode bootstrap peut être appliquée en utilisant $\hat{t}(s, \mathbf{w}^s; \mathbf{0})$ ou $\hat{t}(s, \mathbf{w}^s; \mathbf{c}^k)$ si l'on se sert d'un modèle de régression linéaire. L'utilisation de la première option est parfois plus commode dans le cas de certains projets. Il en a été ainsi dans le cas de notre étude par simulation, puisque l'utilisation de $\hat{t}(s, \mathbf{w}^s; \mathbf{0})$ nous a permis d'éviter de devoir entrer manuellement les valeurs de $\hat{\mathbf{H}}^{ws}$ pour chaque échantillon sélectionné. Une explication informelle de l'égalité $\hat{t}(s, \mathbf{w}^s; \mathbf{0}) = \hat{t}(s, \mathbf{w}^s; \mathbf{c}^k)$ peut être obtenue en traitant $\hat{\beta}^{ws}$ comme une quantité fixe, ce qui est le cas en réalité sous la distribution bootstrap. La statistique bootstrap $\hat{t}(s, \mathbf{w}^s; \mathbf{c}^k)$ peut donc être interprétée comme une statistique destinée à tester l'hypothèse nulle $\mathbf{H}_0^s: \mathbf{H}\hat{\beta}^{ws} = \mathbf{H}\hat{\beta}^{ws}$ ou, alternativement, $\mathbf{H}_0^s: \mathbf{H}\gamma = \mathbf{0}$, où $\gamma = \hat{\beta}^{ws}$. En supposant encore que $\hat{\beta}^{ws}$ est fixe, nous pouvons récrire notre modèle linéaire $\mathbf{E}_m(\mathbf{y}_k | \mathbf{X}^U) = \mathbf{x}_k^t \hat{\gamma}$ sous la forme $\mathbf{E}_m(e_k | \mathbf{X}^U) = \mathbf{x}_k^t \gamma$. Ces observations semblent impliquer que l'utilisation de la statistique bootstrap $\hat{t}(s, \mathbf{w}^s; \mathbf{0})$ équivaut à utiliser $\hat{t}(s, \mathbf{w}^s; \mathbf{c}^k)$, ce qui est en effet vrai.

5. Certaines autres méthodes pour la régression linéaire

Remarque 3 : Nous avons déjà mentionné que l'instruction WEIGHT était nécessaire pour obtenir une statistique pondérée si la méthode de test bootstrap proposée est appliquée en utilisant la procédure REG de SAS. De plus, l'instruction TEST est nécessaire pour demander que la statistique souhaitée soit produite, ainsi que l'instruction « ODS OUTPUT TESTANOVA = » pour sauvegarder ces statistiques demandées dans un ensemble de données SAS spécifié par l'utilisateur.

À la présente section, nous décrivons brièvement certaines méthodes de test dans le contexte de la régression linéaire exposé à la section 4, à savoir deux méthodes naïves qui sont parfois utilisées en pratique, ainsi que des applications spécifiques des méthodes de Rao-Scott, Wald et Bonferroni. Elles seront toutes évaluées dans une étude par simulation à la section 6.

Les méthodes de Bonferroni, de Wald et de Rao-Scott, décrites aux sections 5.2, 5.3 et 5.4 respectivement, nécessitent toutes un estimateur convergent sous m et p de $\hat{\mathbf{V}}^{mp}(\hat{\beta}^{ws})$. Dans l'étude par simulation de la section 6, nous avons utilisé l'estimateur de variance bootstrap

L'hypothèse nulle est rejetée si la valeur est inférieure au seuil de signification α (par exemple, 5 %).

Notons que la statistique qui est soumise au bootstrap est $\hat{f}(s, w_s^*; H\hat{f}_{ws}^*)$ et non $\hat{f}(s, w_s^*; c)$. L'utilisation de

cette dernière ne refléterait pas convenablement la distribution sous l'hypothèse nulle et violerait donc la première ligne directrice donnée dans Hall et Wilson (1991).

Si $\hat{f}(s, w_s; c)$ est un pivot, alors la deuxième ligne directrice de Hall et Wilson (1991) est également satisfaite. Le fait que $t(U; c)$ est asymptotiquement un pivot facilite

certainement l'obtention d'une meilleure procédure de test bootstrap. Cela ne garantit malheureusement pas que $\hat{f}(s, w_s; c)$ sera également asymptotiquement un pivot,

Néanmoins, ne pas utiliser une statistique pivot ne rend pas la procédure de test susmentionnée non valide et ne réduit

pas nécessairement sa puissance. Toutefois, cela pourrait réduire le niveau d'exactitude du test. Comme l'ont souligné

Hall et Wilson (1991), il convient parfois de ne pas tenir compte de la deuxième ligne directrice. Le principal

avantage de l'utilisation de la simple statistique (éventuellement non pivot) $\hat{f}(s, w_s; c)$ dans (2.3) et la statistique

bootstrap $\hat{f}(s, w_s^*; H\hat{f}_{ws}^*)$ tient au fait que, une fois que les poids bootstrap ont été fournis dans le fichier de micro-

données, ces statistiques peuvent être calculées facilement au moyen de logiciels classiques qui ne tiennent pas

compte des caractéristiques du plan d'échantillonnage. En outre, nous montrons à la section 5, au moyen d'une étude

par simulation, que notre méthode de test bootstrap donne des résultats comparables à la méthode de Rao-Scott et de

meilleurs résultats que celles de Wald et de Bonferroni.

4. Un exemple de régression linéaire

Afin de mieux illustrer la théorie dans un contexte pratique, supposons que, sachant X_U , les variables aléatoires y_k , pour $k \in U$, suivent indépendamment une distribution de moyenne $E(y_k | X_U) = x_k' \beta$ et de variance $V(y_k | X_U) = \theta$, où x_k est un vecteur de dimension r de variables linéairement indépendantes pour l'unité k . Rappelons que nous désirons tester l'hypothèse nulle $H_0: H\beta = c$ contre l'hypothèse alternative $H_1: H\beta \neq c$. Si nous pouvons observer la population entière, nous pourrions utiliser la statistique courante

$$t(U; c) = \frac{\hat{\theta}^{1/2} \left(H\hat{f}_U - c \right)' \left(H \left(\sum_{k \in U} x_k x_k' \right)^{-1} H' \right)^{-1} \left(H\hat{f}_U - c \right)}{(4.1)}$$

où

$$\hat{f}_U = \left(\sum_{k \in U} x_k x_k' \right)^{-1} \sum_{k \in U} x_k y_k$$

et

$$\hat{\theta}_U = \frac{\sum_{k \in U} (y_k - x_k' \hat{f}_U)^2}{N - r}$$

Dans (4.1), la statistique $t(U; c)$ suit la distribution $F_{Q, N-r}$ sous l'hypothèse nulle. Elle se réduit à (2.2) quand $\hat{\theta} = r = H = x_k = 1$ dans (4.1).

Une version d'échantillon pondérée de (4.1), qui peut s'écrire sous la forme de (2.3), est

$$t(s, w_s; c) = \frac{\hat{\theta}_{ws}^{1/2} \left(H\hat{f}_{ws} - c \right)' \left(H \left(\sum_{k \in s} w_k x_k x_k' \right)^{-1} H' \right)^{-1} \left(H\hat{f}_{ws} - c \right)}{\hat{\theta}_{ws}^{1/2} \{N - (n - r)/(n - r)\}}, \quad (4.2)$$

où

$$\hat{f}_{ws} = \left(\sum_{k \in s} w_k x_k x_k' \right)^{-1} \sum_{k \in s} w_k x_k y_k \quad (4.3)$$

et

$$\hat{\theta}_{ws} = \frac{\sum_{k \in s} w_k (y_k - x_k' \hat{f}_{ws})^2}{N - r}. \quad (4.4)$$

Par exemple, dans (4.2) la statistique $\hat{f}(s, w_s; c)$

dans la procédure REG de SAS, à condition que $w_k > 0$, pour $k \in s$. Notons qu'elle satisfait l'hypothèse 2 et ne dépend pas de la façon dont les poids sont mis à l'échelle.

De nouveau, si les poids sont rééchantillonnés de manière que $\sum_{k \in s} w_k = n$, le facteur $(N - r)/(n - r)$ disparaît dans (4.2). La statistique de test (4.2) se réduit à (2.4) quand

$\hat{\theta} = r = H = x_k = 1$ dans (4.2), (4.3) et (4.4). La statistique bootstrap $\hat{f}(s, w_s^*; H\hat{f}_{ws}^*)$, ainsi que \hat{f}_{ws}^* et $\hat{\theta}_{ws}^*$, s'obtiennent de manière comparable à $\hat{f}(s, w_s; c)$, \hat{f}_{ws} et $\hat{\theta}_{ws}$ dans (4.2), (4.3) et (4.4) respectivement, excepté

que w_k est remplacé par w_k^* et c est remplacé par $H\hat{f}_{ws}^*$.

Remarque 1 : Notons que w_k^* sera vraisemblablement nul pour certaines unités $k \in s$ (voir, par exemple, Rao et coll. 1992). Dans certains logiciels tels que SAS, le nombre d'observations utilisées dans l'analyse de la b^e réplique bootstrap, n_{b^e} , est égal au nombre d'unités $k \in s$ pour lesquelles $w_k^* > 0$. Ces logiciels peuvent utiliser $n_{b^e} - r$ au lieu de $n - r$ pour calculer la statistique bootstrap $\hat{f}(s, w_s^*; H\hat{f}_{ws}^*)$. Il faut s'assurer que $n - r$ est utilisé et, dans la négative, que la statistique bootstrap calculée au

moyen de ces logiciels est rajustée correctement avant d'appliquer la méthode de test bootstrap proposée. Un moyen d'éviter ce problème consiste à ajouter une très petite valeur positive (par exemple 1×10^{-10}) à chaque poids bootstrap w_k^* pour $k \in s$, afin qu'aucune observation ne

soit exclue du calcul de $\hat{f}(s, w_s^*; H\hat{f}_{ws}^*)$.

L'hypothèse 5 nécessite un commentaire. Cette hypothèse équivaut à exiger que la variance bootstrap $V(\hat{\beta}^{ws})$ soit convergente sous m et p pour

$$(3.1) \quad V^{mp}(\hat{\beta}^{ws}) = E^m V^p(\hat{\beta}^{ws}) + V^m E^p(\hat{\beta}^{ws}).$$

Autrement dit, la distribution bootstrap doit refléter la variabilité due au modèle ainsi qu'au plan d'échantillonnage. Malheureusement, les méthodes bootstrap fondées sur le plan standard ne reflètent que la variabilité due au plan d'échantillonnage, de sorte qu'elles ne déprisent que le premier terme du deuxième membre de (3.1). Donc, elles ne satisfont pas l'hypothèse 5 en général. Toutefois, si la fraction globale d'échantillonnage n/N est négligeable, le deuxième terme du deuxième membre de (3.1) devient lui-même négligeable (par exemple, voir Binder et Roberts 2003), de sorte que l'approximation $V^{mp}(\hat{\beta}^{ws}) \approx E^m V^p(\hat{\beta}^{ws})$ est appropriée et que les méthodes bootstrap fondées sur le plan peuvent être appliquées. Dans le cas de nombreuses enquêtes-ménages, la fraction globale d'échantillonnage est effectivement assez faible. En effet, les poids bootstrap sont souvent obtenus sous l'hypothèse que les fractions d'échantillonnage de premier degré sont faibles (par exemple, Rao et coll. 1992). L'élaboration de méthodes bootstrap qui reflètent les deux termes de (3.1) est un domaine sur lequel porteront de futurs travaux de recherche.

Sous les hypothèses 3 et 4, nous obtenons un deuxième résultat :

Résultat 2 : $\hat{f}(s, w_s^* ; H\hat{\beta}^{ws}) \xrightarrow{*} \sum_{q=1}^Q \hat{\gamma}_q \Omega_q^*$, où $\hat{\gamma}_q =$ pour $q = 1, \dots, Q$, sont les valeurs propres de $\hat{A} = [N\hat{A}(s, w_s^*)]^{-1} [H\hat{Z}(H\hat{\beta}^{ws})]$ et les Ω_q^* sont de nouveau des variables aléatoires khi-carré indépendantes possédant un degré de liberté.

Nous omettons la preuve du résultat 2, car elle est fort semblable à celle du résultat 1 présentée en annexe. À partir des hypothèses 2 et 5, \hat{A} est convergent sous m et p pour A . Donc, en utilisant les résultats 1 et 2, la distribution bootstrap de $\hat{f}(s, w_s^* ; H\hat{\beta}^{ws})$ est asymptotiquement la même que la distribution sous m et p de $\hat{f}(s, w_s^* ; H\hat{\beta})$, qui est elle-même identique à la distribution sous m et p de $\hat{f}(s, w_s^* ; c)$ sous l'hypothèse nulle, c'est-à-dire celle que nous voulons approximer. Cela suggère la procédure de test bootstrap suivante :

- i) obtenir les poids bootstrap, w_s^k , pour $k \in s$ et $b = 1, \dots, B$;
- ii) calculer $\hat{f}(s, w_s^* ; H\hat{\beta}^{ws})$, pour $b = 1, \dots, B$;
- iii) puisqu'une grande valeur de $\hat{f}(s, w_s^* ; c)$ donne lieu au rejet de l'hypothèse nulle, calculer le seuil de signification observé (valeur p) selon
$$\frac{B}{\#\{\hat{f}(s, w_s^* ; H\hat{\beta}^{ws}) > \hat{f}(s, w_s^* ; c)\}}.$$

Sous l'hypothèse nulle, les deux derniers termes du deuxième membre de (2.6) disparaissent et nous avons $\hat{f}(s, w_s^* ; c) = \hat{f}(s, w_s^* ; H\hat{\beta})$. Si l'hypothèse nulle est fausse, le troisième terme du deuxième membre de (2.6) domine les autres à mesure que la taille d'échantillon augmente, puisque les premier, deuxième et troisième termes sont d'ordre $O_p(1)$, $O_p(\sqrt{n})$ et $O_p(n)$ respectivement, à condition que les hypothèses 1 et 2 soient vérifiées. En outre, puisque $\hat{A}(s, w_s^*)$ est définie positive, le troisième terme est toujours positif. Par conséquent, l'observation d'une grande valeur positive de $\hat{f}(s, w_s^* ; c)$ comparativement à un centile élevé de la distribution de $\hat{f}(s, w_s^* ; H\hat{\beta})$ est une indication que l'hypothèse nulle pourrait être fausse.

3. La méthode bootstrap proposée

Soit w_s^* un poids bootstrap aléatoire pour l'unité k , Rao et coll. (1992), et soit w_s^j le vecteur de dimension n qui contient le poids bootstrap aléatoire w_s^k dans son k^e élément. L'estimateur par le bootstrap $\hat{\beta}^{ws}$ s'obtient de la même manière que $\hat{\beta}^{ws}$ en remplaçant le poids de sondage w_i par sa version bootstrap w_i^* pour chaque unité échantillonnée. Nous désignons aussi par w_s^b , pour $b = 1, \dots, B$, les B vecteurs de dimension n contenant les poids bootstrap w_s^k dans leur k^e élément. Ces B vecteurs sont tirés indépendamment et ont la même distribution que w_s^* ; celle-ci est appelée distribution bootstrap et est désignée par le symbole « * ». Le b^e estimateur par le bootstrap $\hat{\beta}^{ws_b}$ est défini d'une manière évidente.

Avant de décrire notre procédure de test bootstrap, nous commençons par introduire trois hypothèses supplémentaires relatives à la construction des poids bootstrap :

Hypothèse 3 : $\sqrt{n}(\hat{\beta}^{ws_b} - \beta^{ws}) \xrightarrow{*} N(0, \Sigma)$, où $\xrightarrow{*}$ désigne la convergence en distribution bootstrap et Σ est la matrice de variance-covariance bootstrap asymptotique de $\sqrt{n}\hat{\beta}^{ws}$.

Hypothèse 4 : $n\hat{A}(s, w_s^*)$ est convergent en distribution bootstrap pour $n\hat{A}(s, w_s^*)$.

Hypothèse 5 : $\hat{\Sigma}$ est convergent sous m et p pour Σ .

Les hypothèses 3 et 4 sont les analogues bootstrap des hypothèses 1 et 2 et devraient être satisfaites avec la plupart des méthodes bootstrap (par exemple, celles décrites dans l'article de synthèse de Sitter 1992) et modèles (par exemple, modèle de régression linéaire, modèle de régression logistique). Pour des renseignements plus détaillés, le lecteur est invité à consulter Shao et Tu (1995, chapitre 6 ; en particulier la section 6.4.4).

$$\hat{f}(s, w_s; c) = (H\beta_{ws} - c)' \{ \hat{A}(s, w_s) \}^{-1} (H\beta_{ws} - c). \quad (2.3)$$

Le vecteur w_s de dimension n contient le poids de sondage de l'unité échantillonnée k dans son k^e élément, désigné par w_k, β_{ws} est un estimateur pondéré de β et $\hat{A}(s, w_s)$ est un analogue pondéré de $A(s)$ en ce sens

que chaque unité échantillonnée k est pondérée par son poids de sondage w_k , alors qu'aucune pondération n'a lieu avec $A(s)$. Nous avons donc $\hat{A}(s, I_s) = A(s)$, où I_s est un vecteur échantillon de valeurs un. Par conséquent, la statistique $\hat{f}(s, w_s; c)$ est aussi un analogue pondéré de $f(s; c)$ et nous avons $f(s, I_s; c) = f(s; c)$. Si la statistique $\hat{f}(s; c)$ peut être calculée en se servant d'un projetiel classique qui n'a pas nécessairement été développé pour traiter des données d'enquête, la statistique $\hat{f}(s, w_s; c)$ peut aussi être calculée en utilisant le même projetiel, à condition que ce dernier permette la pondération de chaque observation par son poids de sondage.

Ordinairement, le poids de sondage w_k pour une unité de sondage sont construits de manière que les deux hypothèses suivantes soient vérifiées :

$$\text{Hypothèse 1 : } \sqrt{n}(\beta_{ws} - \beta) \xrightarrow{mp} N(0, \Sigma), \text{ où}$$

\xrightarrow{mp} désigne la convergence en distribution sous le modèle et le plan d'échantillonnage, et Σ est la matrice de variance-covariance asymptotique de $\sqrt{n}\beta_{ws}$ sous le

plan d'échantillonnage probabiliste.

Hypothèse 2 : $n\hat{A}(s, w_s)$ est symétrique, définie positive et convergente sous m et p pour une matrice de mise à l'échelle fixe, définie positive et symétrique, \hat{A} .

Notons que l'hypothèse 2 n'exige pas que $\hat{A}(s, w_s)$ soit convergente sous p pour $A(U)$. En effet, $NA(U)$ sera habituellement convergente sous m pour A . D'autres choix de la matrice de mise à l'échelle $\hat{A}(s, w_s)$ dans (2.3) sont possibles. Par exemple, elle pourrait être remplacée par un estimateur de la variance par rapport au plan de $H\beta_{ws}$ sous échantillonnage aléatoire simple (par exemple, Rao et Scott 1981). Un autre choix est la statistique de Wald courante. On l'obtient en remplaçant $\hat{A}(s, w_s)$ dans (2.3) par $V_{mp}^{ws}(H\beta_{ws})$, qui est un estimateur convergent sous m et p de $V_{mp}^{ws}(H\beta_{ws})$, la variance de $H\beta_{ws}$ évaluée par rapport au modèle et au plan d'échantillonnage. Comme nous le mentionnons au paragraphe qui suit l'équation (2.3), l'un des avantages de l'utilisation d'une matrice de mise à l'échelle $\hat{A}(s, w_s)$ telle que $\hat{A}(s, I_s) = A(s)$ est que la statistique de test résultante $\hat{f}(s, w_s; c)$ peut alors être calculée directement en utilisant des projectiels classiques, à

condition qu'ils permettent la pondération de chaque observation par son poids de sondage, ce qui est plus commode pour les utilisateurs de données d'enquête. Poursuivant l'exemple susmentionné, nous pouvons définir notre statistique de test pondérée sous la forme

$$\hat{f}(s, w_s; c) = \frac{\{\hat{N} - 1\}/(n - 1) \{ \hat{\theta}^{ws}/N \}}{(\beta_{ws} - c)^2}. \quad (2.4)$$

où $\hat{N} = \sum_{k \in s} w_k, \hat{\beta}_{ws} = \sum_{k \in s} w_k y_k / \sum_{k \in s} w_k$ et $\hat{\theta}^{ws} = \sum_{k \in s} w_k (y_k - \beta_{ws})^2 / \hat{N} - 1$. Dans (2.4), la matrice de mise à l'échelle pondérée sous-jacente est $\hat{A}(s, w_s) = \{ \{ \hat{N} - 1 \} / (n - 1) \} \{ \hat{\theta}^{ws} / N \}$, qui ne dépend pas de la façon dont les poids sont rééchelonnés. S'ils le sont de manière que $\sum_{k \in s} w_k = n$, ce que font habituellement les analystes, le facteur $\{ \hat{N} - 1 \} / (n - 1)$ disparaît. Le rôle de ce facteur, conjugué à d'autres conditions de régularité, est de satisfaire l'hypothèse 2. Si l'on choisit d'utiliser le système SAS[®], la statistique de test (2.4) s'obtient en utilisant l'instruction WEIGHT dans les procédures standard. Si l'hypothèse nulle est vérifiée, il est bien connu que (2.4) ne suit malheureusement pas la distribution χ^2_{n-1} ou $F_{1, n-1}$ sous le modèle et le plan d'échantillonnage.

Pour obtenir une procédure de test valide, nous devons approximer la distribution de $\hat{f}(s, w_s; c)$ sous l'hypothèse nulle. Nous pouvons pour cela utiliser le résultat suivant :

$$\text{Résultat 1 : } \hat{f}(s, w_s; c) \xrightarrow{mp} \sum_{q=1}^g \lambda_q \Omega_q, \text{ où } \lambda_q$$

Ω_q sont des variables aléatoires kh- $(\hat{A}^{-1})'(\Sigma H)$, et Ω_q sont les valeurs propres de $\hat{A} = \hat{Q}, \hat{Q}$, pour $q = 1, \dots, g$, sont les valeurs propres de \hat{A} vérifiée (c'est-à-dire $H\beta$ est nul, nous avons donc

$$\hat{f}(s, w_s; c) \xrightarrow{mp} \sum_{q=1}^g \lambda_q \Omega_q. \quad (2.5)$$

Rao et Scott (1981) ont utilisé un résultat comparable pour construire leur procédure de test. Ils ont approximé une distribution telle que (2.5) par une distribution du khi-carré mise à l'échelle qui concorde avec les deux premiers moments estimés du deuxième membre de (2.5). Au lieu de cela, nous approximations la distribution de $\hat{f}(s, w_s; c)$ sous l'hypothèse nulle en utilisant des poids bootstrap, ce que nous décrivons à la section suivante.

Avant de fournir des détails sur notre procédure de test, il convient de souligner que, dans (2.3), $\hat{f}(s, w_s; c)$ peut s'écrire

$$\hat{f}(s, w_s; c) = \hat{f}(s, w_s; H\beta) + 2(H\beta_{ws} - H\beta)' \{ \hat{A}(s, w_s) \}^{-1} (H\beta - c) + (H\beta - c)' \{ \hat{A}(s, w_s) \}^{-1} (H\beta - c). \quad (2.6)$$

Le problème des tests d'hypothèses à partir des données d'enquête complexes a été bien étudié les 30 dernières années (par exemple Rao et Scott 1981 ; Fay 1985 ; Thomas et Rao 1987 ; Korn et Graubard 1990 ; Korn et Graubard 1991 ; Graubard et Korn 1993 ; Thomas, Singh et Roberts 1996 ; Rao et Thomas 2003). Toutefois, sauf peut-être pour estimer les variances-covariances inconnues intervenant dans ces méthodes, la méthode bootstrap ne semble pas encore avoir été étudiée pour les tests d'hypothèses. L'objectif du présent article est donc de proposer une méthode bootstrap pour les tests d'hypothèses au sujet d'un vecteur de paramètres inconnus d'un modèle quand l'échantillon a été tiré d'une population finie. Le plan d'échantillonnage probabilistes utilisé pour sélectionner l'échantillon peut être informatif ou non. Informellement parlant l'échantillonnage est informatif quand le modèle vérifié pour l'échantillon sélectionné diffère de celui qui tient pour l'ensemble de la population ; sinon, l'échantillonnage n'est pas informatif.

Notre méthode s'appuie sur des statistiques de tests fondées sur un modèle dans lesquelles sont intégrés les poids de sondage. Habituellement, ces statistiques se calculent facilement à l'aide de logiciels classiques. Pour approximer la distribution sous l'hypothèse nulle de ces statistiques pondérées fondées sur un modèle, nous utilisons des poids bootstrap. L'un des avantages de notre méthode bootstrap par rapport aux méthodes existantes de tests d'hypothèses à partir des données d'enquête est que, après avoir reçu les ensembles de poids bootstrap, les analystes peuvent l'appliquer très facilement, même s'ils ne disposent pas de logiciels spécialisés pour le traitement des données d'enquêtes complexes.

À la section 2, nous présentons la notation et le problème. À la section 3, nous décrivons et justifions la méthode bootstrap que nous proposons pour les tests d'hypothèses à partir des données d'enquête. À la section 4, nous donnons un exemple de régression linéaire pour illustrer la théorie. À la section 5, nous décrivons brièvement les méthodes de Rao-Scott (Rao et Scott 1981), de Wald et de Bonferroni, qui sont des méthodes de rechangement, dans le cas de tests d'hypothèses au sujet d'un vecteur des paramètres d'un modèle de régression linéaire. À la section 6, au moyen d'une étude par simulation, nous comparons ces méthodes et les comparons à la méthode bootstrap que nous proposons. Enfin, à la dernière section, nous concluons par un bref résumé et une discussion.

2. Préliminaires

Nous supposons qu'une population finie U de taille N a été créée conformément à un modèle, spécifié par l'analyse, qui décrit la distribution conditionnelle $F(y_U | \mathbf{x}_U; \boldsymbol{\beta}, \boldsymbol{\theta})$. Le vecteur y_U de dimension N

contiennent les valeurs dans la population d'une variable dépendante y , \mathbf{X}_U est une matrice à N lignes qui contient les valeurs dans la population d'un vecteur de variables indépendantes \mathbf{x} , $\boldsymbol{\beta}$ est un vecteur de dimension r des paramètres inconnus du modèle et $\boldsymbol{\theta}$ est un vecteur de paramètres inconnus supplémentaires du modèle. Nous soulignons les tests des hypothèses au sujet de $\boldsymbol{\beta}$, mais nous soulignons également que, si l'on pouvait observer la population complète U , on utiliserait une statistique de test $t(U; \mathbf{c})$ pour tester l'hypothèse linéaire multiple $H_0: \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ contre l'hypothèse alternative $H_1: \mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$. La matrice \mathbf{H} de dimensions $\bar{Q} \times r$ est utilisée pour définir l'hypothèse à tester et \mathbf{c} est un vecteur de dimension \bar{Q} de constantes spécifiées par l'analyste. Idéalement, $t(U; \mathbf{c})$ est asymptotiquement un pivot ; autrement dit, elle suit une distribution asymptotique qui ne dépend d'aucun paramètre inconnu. Nous considérons les statistiques qui ont la forme quadratique suivante :

$$t(U; \mathbf{c}) = (\mathbf{H}\boldsymbol{\beta}_U - \mathbf{c})' \mathbf{A}(U)^{-1} (\mathbf{H}\boldsymbol{\beta}_U - \mathbf{c}), \quad (2.1)$$

où $\boldsymbol{\beta}_U$ est un estimateur convergent de $\boldsymbol{\beta}$ sous le modèle et $\mathbf{A}(U)$ est une matrice de mise à l'échelle. Habituellement, $\mathbf{A}(U)$ est symétrique et définie positive.

À titre d'exemple, supposons que les y_i pour toutes les unités de la population $k \in U$, sont des variables aléatoires indépendantes et identiquement distribuées de moyenne β et de variance θ , et que nous souhaitons tester l'hypothèse nulle $H_0: \beta = c$. Dans cet exemple, $\bar{Q} = 1$, $r = 1$, $\mathbf{H} = 1$ et $\mathbf{X}_U = \mathbf{1}_U$, où $\mathbf{1}_U$ est un vecteur de population de valeurs un. Une statistique de test courante pour ce problème est

$$t(U; c) = \frac{(\bar{y}_U - c)^2}{\hat{\theta}_U / N}, \quad (2.2)$$

où $\bar{y}_U = \sum_{k \in U} y_k / N$ et $\hat{\theta}_U = \sum_{k \in U} (y_k - \bar{y}_U)^2 / (N - 1)$. La statistique (2.2) est de la même forme que (2.1) si nous posons que $\mathbf{A}(U) = \hat{\theta}_U / N$. On suppose habituellement que cette statistique suit la distribution χ^2_{r-1} ou $F_{r-1, N-1}$ sous l'hypothèse nulle.

Comme cela est ordinairement le cas, un échantillon aléatoire s de taille n est tiré de la population finie U conformément à un plan d'échantillonnage probabiliste donné $p(s)$. Puisque la variable dépendante y et, éventuellement, les variables indépendantes \mathbf{x} ne sont pas observées pour les unités non échantillonnées, il pourrait être préférable d'utiliser la statistique $t(s; \mathbf{c})$ plutôt que $t(U; \mathbf{c})$. Dans l'exemple susmentionné, cela mènerait à $t(s; \mathbf{c}) = n(\bar{y}_s - c)^2 / \hat{\theta}_s$, où $\bar{y}_s = \sum_{k \in s} y_k / n$ et $\hat{\theta}_s = \sum_{k \in s} (y_k - \bar{y}_s)^2 / (n - 1)$. Cependant, si l'échantillonnage est approprié, ce qui est indubitablement plus fréquent, d'utiliser une statistique de test pondérée de la forme

Une méthode bootstrap pratique pour les tests d'hypothèses à partir des données d'enquête

Jean-François Beaumont et Cynthia Bocci¹

Résumé

Le recours à la méthode bootstrap est de plus en plus répandu dans le contexte des enquêtes par sondage réalisées par les organismes statistiques nationaux. Dans la plupart des applications, plusieurs ensembles de poids bootstrap sont fournis aux analystes avec le fichier de microdonnées d'enquête. Jusqu'à présent, l'utilisation de la méthode en pratique semble avoir été limitée principalement aux problèmes d'estimation de la variance. Dans le présent article, nous proposons une méthode bootstrap pour les tests d'hypothèses au sujet d'un vecteur de paramètres inconnus d'un modèle quand l'échantillon a été tiré d'une population finie. Le plan d'échantillonnage probabiliste utilisé pour sélectionner l'échantillon peut être informatif ou non. Notre méthode s'appuie sur des statistiques de test fondées sur un modèle dans lesquelles les poids de bootstrap pour les tests d'hypothèses sont habituellement faciles à calculer en se servant de propriétés statistiques classiques. Nous approximations la distribution sous l'hypothèse nulle de ces statistiques pondérées fondées sur un modèle en utilisant des poids bootstrap. L'un des avantages de notre méthode bootstrap par rapport aux méthodes existantes de test d'hypothèses à partir des données d'enquête est qu'après avoir reçu les ensembles de poids bootstrap, les analystes peuvent l'appliquer très facilement, même s'ils ne disposent pas de logiciels spécialisés pour le traitement des données d'enquêtes complexes. En outre, nos résultats de simulation laissent entendre que, dans l'ensemble, la méthode donne des résultats comparables à ceux de la méthode de Rao-Scott et meilleurs que ceux des méthodes de Wald et de Bonferroni quand on teste des hypothèses au sujet d'un vecteur de paramètres d'un modèle de régression linéaire.

Mots clés : Poids bootstrap ; analyse de données d'enquête ; test d'hypothèses ; échantillonnage informatif ; régression linéaire ; paramètres d'un modèle.

1. Introduction

La méthode bootstrap est utilisée de plus en plus fréquemment dans le contexte des enquêtes par sondage réalisées par les organismes statistiques nationaux. Les principales raisons semblent être qu'elle permet de traiter plusieurs situations qu'il serait difficile de résoudre autrement (par exemple ajustement des poids pour la non-réponse, calage, statistiques non lissées) et qu'elle est commode pour les analystes. Dans la plupart des applications, plusieurs ensembles de poids bootstrap sont fournis aux analystes avec le fichier de microdonnées d'enquête ; aucune autre information sur le plan n'est donnée. Ces poids sont habituellement obtenus en supposant que les fractions d'échantillonnage de premier degré sont suffisamment faibles pour qu'un plan d'échantillonnage sans remise puisse être approximé correctement par un plan d'échantillonnage avec remise. Le lecteur trouvera dans Rao, Wu et Yue (1992) une description succincte, mais claire, d'une méthode de construction de poids bootstrap sous cette hypothèse dans le cas d'un plan d'échantillonnage stratifié à plusieurs degrés.

Juste ici, l'utilisation de la méthode en pratique semble avoir été limitée principalement aux problèmes d'estimation de la variance (par exemple, Langlet, Faucher et Lesage

2003 ; Yeo, Manel et Lu 1999 ; Hughes et Brodsky 1994).

Dans le domaine de la recherche, les travaux ont visé principalement à découvrir une méthode bootstrap appropriée pour estimer la variance quand l'échantillon est tiré sans remise d'une population finie (voir Sitter 1992 ou Shao et Tu 1995, chapitre 6, pour une revue des méthodes). Certains auteurs ont également étudié le problème de la détermination des intervalles de confiance bootstrap pour un paramètre de population finie (par exemple, Rao et Wu 1988 ; Kovar, Rao et Wu 1988 ; Sitter 1992 ; Rao et coll. 1992). Autant que nous sachions, il ne semble exister aucune publication sur les tests d'hypothèses utilisant la méthode bootstrap dans le cas des sondages, quoique le problème ait été étudié dans le contexte de la statistique classique. Le lecteur trouvera dans Hall et Wilson (1991) une discussion des tests d'hypothèses bootstrap et dans Efron et Tibshirani (1993), un excellent compte rendu de la méthode bootstrap en statistique classique. Les travaux de Graubard, Korn et Midlune (1997), qui ont appliqué la méthode bootstrap paramétrique classique à des données d'enquête pour tester l'ajustement d'un modèle de régression logistique, méritent également d'être soulignés. Leur procédure est valide quand l'échantillonnage n'est pas informatif.

1. Jean-François Beaumont, Statistique Canada, Division de la recherche et de l'innovation en statistique, Pré Tunney, immeuble R-H-Cats, 16^e étage, Ottawa (Ontario), Canada, K1A 0T6. Courriel : Jean-Francois.Beaumont@statcan.gc.ca ; Cynthia Bocci, Statistique Canada, K1A 0T6. Courriel : Jean-Francois.Bocci@statcan.gc.ca

- Chambers, R.L., et Skinner, C.J. (2003). *Analysis of Survey Data*. New York : John Wiley & Sons, Inc.
- Choudhry, G. (2000). The 1998 Survey of Mental Health Organizations Survey Design. Westat technical report prepared for Center for Mental Health Services, Substance Abuse and Mental Health Services Administration (SAMHSA), disponible sur demande au SAMHSA.
- Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.
- Deville, J.-C., et Sarda, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Duchesne, P. (1999). Estimateurs de calage robuste. *Techniques d'enquête*, 25, 47-60.
- Gwet, J., et Rivest, L. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- Hoaglin, D.C., et Welsh, R.E. (1978). The hat matrix in regression and ANOVA (Corr. 78V32 p146). *The American Statistician*, 32, 17-22.
- Hulliger, B. (1995). Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes. *Techniques d'enquête*, 21, 89-97.
- Hurwich, C.M., et Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.
- Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York : John Wiley & Sons, Inc.
- Li, J. (2007). *Regression Diagnostics for Complex Survey Data: Identification of Influential Observations*. Thèse de doctorat non-publiée, Université of Maryland.
- Li, J., et Valliant, R. (2006). Influence analysis in linear regression with sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3330-3337.
- Manderscheid, R.W., et Henderson, M.J. (2002). Mental Health, Rockville MD USA : Substance Abuse and Mental Health Services Administration. Disponible au <http://www.mentalhealth.samhsa.gov/publications/allpubs/SMA04-3938/appendixA.asp>
- Moreno-Rebollo, J.L., Muñoz-Reyes, A., et Muñoz-Pichardo, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.
- Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.
- Potter, F.J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.
- Potter, F.J. (1993). The effect of weight trimming on nonlinear survey estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.
- Skinner, C.J. (2003). Introduction to Part B, Chapitre 6 dans *Analysis of Survey Data*, (Eds. R. Chambers et C. Skinner). New York : John Wiley & Sons, Inc.
- Skinner, C.J., Holt, D., et Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. New York : John Wiley & Sons, Inc.
- Smith, T.M.F. (1989). Introduction to Part B, Chapitre 6 dans *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt et T.M.F. Smith). New York : John Wiley & Sons, Inc.
- Valliant, R., Dorfman, A.H., et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York : John Wiley & Sons, Inc.
- Weisberg, S. (2005). *Applied Linear Regression*, Troisième édition. New York : John Wiley & Sons, Inc.
- Welsh, A.H., et Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B, Methodological*, 60, 413-428.
- Zaslavsky, A.M., Schenker, N., et Belin, T.R. (2001). Downweighting influential clusters in surveys: Application to the 1990 post enumeration Survey. *Journal of the American Statistical Association*, 96, 858-869.
- Zhang, P. (1992). Influence after variable selection in linear regression models. *Biometrika*, 79, 741-746.

des intervalles de confiance dont la couverture est inférieure au taux nominal et à des taux d'erreur de type I exagérés dans les tests de vérification d'hypothèses (Li 2007). Ce phénomène est semblable aux problèmes bien connus qui se posent dans la régression séquentielle (Hurvich et Tsai 1990, Zhang 1992). Donc, un sujet de recherche intéressant semble être l'élaboration de procédures d'inférence pour la construction d'intervalles de confiance et l'exécution de tests de vérification d'hypothèse qui tiennent compte des effets de l'élimination ou de la sous-pondération de certains points.

Pour les données d'enquête complexes, la matrice chapeau ne fait intervenir aucune caractéristique du plan de sondage, sauf les poids de sondages et peut être utilisée pour détecter les cas dont les poids ou les valeurs des variables explicatives sont atypiques. D'autres statistiques diagnostiques, comme la distance de Cook, contiennent des estimations de la variance qui doivent tenir compte des caractéristiques du plan d'échantillonnage complexe, telles que la stratification et la mise en grappes. L'adaptation et l'extension d'approches diagnostiques supplémentaires à l'analyse des données d'enquête seront étudiées dans l'avenir.

7. Remerciements

Le présent article est fondé sur des travaux financés par la National Science Foundation des États-Unis en vertu de la subvention n° 0617081. Toutes les opinions, constatations et conclusions ou recommandations exprimées dans l'article sont celles des auteurs et ne reflètent pas forcément celles de la National Science Foundation. Nous remercions le rédacteur en chef et les examinateurs de leurs commentaires constructifs, qui nous ont permis d'améliorer considérablement l'article.

Annexe

Inclusion des unités sélectionnées avec certitude dans l'estimation de l'erreur-type

Dans l'étude empirique décrite à la section 5, nous avons inclus les unités sélectionnées avec certitude dans les calculs de l'erreur-type. Nous esquissons ici la justification de cette façon de procéder. Sous le modèle général (3), la variance sous le modèle de $\beta = (X_T^T W X_T)^{-1} X_T^T W Y$, l'estimateur utilisé dans l'étude empirique, est $\text{var}_M(\beta) = A^{-1} X_T^T W W X A^{-1} \sigma^2$ où $A = X^T W X$ et $V = \text{diag}(v_i)_{i \in S}$. L'estimateur sandwich de variance utilisé dans l'étude présentée à la section 5 est défini comme étant

$$v(\beta) = A^{-1} \frac{n-1}{n} \sum_{i \in S} (z_i - \bar{z})(z_i - \bar{z})^T A^{-1} \quad (4)$$

où $z_i = w_i e_i x_i$ avec $e_i = Y_i - x_i \beta$ et $\bar{z} = \sum_{i \in S} w_i e_i x_i / n$. Cet estimateur est convergent par rapport au plan (voir Binder 1983) dans l'échantillonnage à un seul degré si les unités sont échantillonnées avec remise avec probabilité égale à w_i^{-1} , et qu'aucune unité n'est sélectionnée avec certitude. Si l'échantillon contient des unités sélectionnées avec certitude, la formule de $v(\beta)$ doit être modifiée pour estimer la variance par rapport au plan : les unités sélectionnées avec certitude doivent être exclues des sommes figurant dans (4) et de \bar{z} , et n doit être remplacé par n_{nc} , le nombre d'unités sélectionnées sans certitude. Dans le cas extrême d'un recensement, l'estimateur de variance par rapport au plan se réduirait à zéro.

L'estimateur donné par (4) est approximativement sans biais par rapport au modèle sous (3), que l'échantillon contienne ou non des unités sélectionnées avec certitude. Dans (4), la matrice du milieu peut être développée sous la forme $\sum_{i \in S} (z_i - \bar{z})(z_i - \bar{z})^T = \sum_{i \in S} z_i z_i^T - n \bar{z} \bar{z}^T$. En supposant que $e_i \approx Y_i - x_i^T \beta$, l'espérance par rapport au modèle sous (3) du premier terme est $E_M(\sum_{i \in S} z_i z_i^T) = X^T W W X \sigma^2$, tandis que $E_M(n \bar{z} \bar{z}^T) = n^{-1} X^T W W X \sigma^2$. La substitution de ces espérances donne $E_M[v(\beta)] = \text{var}_M(\beta)$, expression qui est vérifiée même si certaines unités sont sélectionnées avec certitude. Cela montre aussi que $v(\beta)$ est robuste au sens où il reflète correctement la contribution des variances hétéroscédastiques dans (3) à la variance par rapport au modèle de β , même si V peut être inconnue et non prise en compte dans l'estimation de β .

Bibliographie

Beaumont, J.-F., et Alavi, A. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 217-231.

Belsey, D.A., Kuh, E. et Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York : John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A., et Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters, Chapitre 3 dans *Analysis of Survey Data*, (Éds. R. Chambers et C. Skinner). New York : John Wiley & Sons, Inc.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Chambers, R.L., Dorfman, A.H. et Sverchkov, M.Y. (2003). Nonparametric regression with complex survey data, Chapitre 11 dans *Analysis of Survey Data*, (Éds. R. Chambers et C. Skinner). New York : John Wiley & Sons, Inc.

Tableau 5
Ratios des estimations des paramètres par MCO et par PPS avant et après l'élimination des observations dont l'effet de levier est supérieur à 0,007 dans la régression pour la SMHO

Ratio des estimations par MCO aux estimations par PPS		
Avec tous les points		
Avec élimination des points à effet de levier élevé		
Lits	1,16	0,91
Patients ajoutés	1,26	0,95

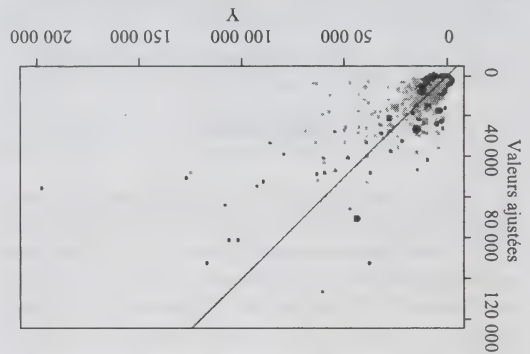
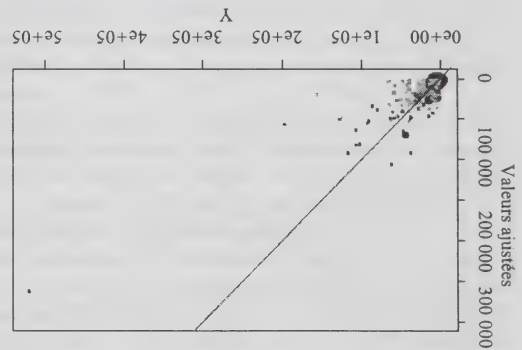


Figure 3 Tracé des valeurs ajustées en fonction des valeurs de Y . La droite de référence est tracée à $Y = X$. Le graphique supérieur comprend tous les points. Dans le graphique inférieur, l'observation extrême 818 est omise. Les points à effet de levier élevé basés sur la PPS sont représentés par des cercles pleins, foncés, dans chaque graphique

Les effets de levier sont habituellement combinés aux résidus pour déterminer quels points sont influents dans l'ajustement du modèle de régression, parce que les résidus peuvent être utilisés pour détecter les valeurs de Y divergentes. Un diagramme de dispersion des valeurs ajustées provenant de la régression sous PPS en fonction des valeurs élevées sont présentes par des cercles pleins foncés. Les

6. Conclusion

Les effets de levier et les résidus sont des composantes essentielles des statistiques diagnostiques destinées à détecter l'influence importante d'une seule observation ou d'un groupe d'observations sur un modèle linéaire ajusté. Les ensembles de données d'enquête peuvent contenir des observations influentes, que l'on soutienne que le plan d'échantillonnage est ignorable et que les moindres carrés ordinaires peuvent être utilisés, ou que le plan doit être pris en compte et les poids de sondage utilisés. Les points qui sont influents dans les deux cas ne sont pas nécessairement les mêmes, comme nous l'avons illustré ici.

Une fois que les points dont l'effet de levier est grand sont identifiés, une question importante est celle de savoir comment les traiter pour l'inférence. Deux options consistent soit à les sous-pondérer, soit à les éliminer entièrement de l'ajustement du modèle. La sous-pondération semble insatisfaisante en général, puisqu'un point peut avoir un effet de levier élevé non pas à cause d'un grand poids, mais plutôt parce qu'il possède une ou plusieurs valeurs X inhabituelles. La sous-pondération peut être raisonnable dans le contexte d'une approche fondée sur un modèle, en supposant que le modèle proprement dit est correctement spécifié. Cependant, la notion fondée sur le plan de sondage d'estimer un paramètre dans des conditions de recensement pourrait alors se perdre. Si un point possède un effet de levier important à cause de valeurs X extrêmes, il est possible qu'il ne suive pas du tout le modèle et qu'il doive être éliminé.

Cependant, utiliser une procédure mécanique qui rejette automatiquement de nombreuses observations influentes estimation de l'erreur-type trop petites, ce qui se traduit par

des changements appréciables dans les estimations des paramètres tant par les MCO que par PPS. Cela illustre aussi qu'un seul point peut avoir une incidence sur les erreurs-types des pentes estimées dans une régression pondérée par les poids de sondage, comme c'est également le cas dans une régression par les MCO. L'observation 818 possède un grand résidu (voir figure 3) ; son omission fait baisser l'erreur-type pour le nombre de lits, de 13,14 dans le tableau 3 à 8,04 dans le tableau 4. Notons que, si l'unité 818 avait un grand poids, son résidu serait vraisemblablement plus petit, puisqu'il aurait plus d'effet sur l'ajustement. Le cas échéant, l'erreur-type pourrait actuellement être plus faible que quand l'unité 818 est incluse.

Tableau 3
Estimation par MCO et PPS des paramètres de la régression pour la SMHO en utilisant la totalité des 875 cas échantillonnés

Variables	Estimation par les MCO		Coefficient E.-T. t	Coefficient E.-T. t
	Estimation par PPS	Estimation par PPS		
Ordonnée	-1 201,73	526,19	-2,28	514,08
Nbre de lits à l'origine	94,16	3,03	31,08	81,23
Nbre de patients ajoutés	2,31	0,13	18,50	1,84
				0,76 2,43

Tableau 4
Estimation par MCO et par PPS des paramètres de la régression pour la SMHO

Variables	Estimation par les MCO		Coefficient E.-T. t	Coefficient E.-T. t
	Estimation par PPS	Estimation par PPS		
Ordonnées	2 987,55	490,54	6,09	1 993,86
Nbre de lits à l'origine	69,27	4,35	15,94	75,82
Nbre de patients ajoutés	0,95	0,20	4,71	1,00
				0,21 4,73
(ii) Suppression de l'observation 818				
Ordonnées	1 979,51	537,93	3,68	2 281,17
Nbre de lits à l'origine	81,80	2,92	27,98	68,69
Nbre de patients ajoutés	1,19	0,14	8,41	0,79
				0,29 2,75

i) Suppression des observations dont l'effet de levier est supérieur à 0,007				
Ordonnées	2 987,55	490,54	6,09	1 993,86
Nbre de lits à l'origine	69,27	4,35	15,94	75,82
Nbre de patients ajoutés	0,95	0,20	4,71	1,00
				0,21 4,73
(ii) Suppression de l'observation 818				
Ordonnées	1 979,51	537,93	3,68	2 281,17
Nbre de lits à l'origine	81,80	2,92	27,98	68,69
Nbre de patients ajoutés	1,19	0,14	8,41	0,79
				0,29 2,75

Un autre point qui se dégage des tableaux 3 et 4 est que les estimations par les MCO et par PPS sont nettement plus proches les unes des autres après l'élimination des points à effet de levier élevé qu'avant. Comme l'illustre le tableau 5, 26 % aux estimations par PPS quand tous les points sont retenus, mais sont inférieures de 9 % et de 5 % à ces estimations après avoir laissé tomber certains points.

Étant donné que certains cas éventuellement influents ont été décelés, l'étape suivante consiste à voir quel effet ils ont sur les estimations des paramètres. Le tableau 3 donne les estimations des paramètres fondées sur les MCO et sur la PPS en utilisant tous les cas. Le tableau 4 donne les estimations fondées sur les MCO et sur la PPS (i) en omettant les cas à effet de levier élevé et (ii) en omettant ceux pour lesquels $h_{ii} > 0,007$. Toutefois, il convient de souligner que les ensembles de points ayant un effet de levier élevé ne sont pas les mêmes dans les régressions par les MCO et par PPS. Les erreurs-types sont estimées au moyen de la formule des MCO habituelle et au moyen de l'estimateur sandwich (Binder 1983) pour les estimations PPS.

Si nous comparons les tableaux 3 et 4, nous constatons que les estimations par les MCO changent considérablement après l'élimination des points à effet de levier élevé (section (i) du tableau 4). L'ordonnée à l'origine de la droite des MCO, qui est significative dans les deux tableaux, passe d'une valeur négative à une valeur positive. La pente de la droite des MCO pour le nombre de lits diminue d'environ 26 % (de 94,16 à 69,27) quand les points à effet de levier élevé sont éliminés. Dans le cas de la pente pour le nombre de patients ajoutés, la diminution est d'environ 59 %. Les estimations PPS pour les nombres de lits et de patients ajoutés sont également sensibles à la présence de points à effet de levier élevé, les pentes diminuant de 7 % et de 46 %, respectivement. Dans tous les cas, les pentes sont significatives, de sorte que la conclusion qualitative selon laquelle les dépenses sont corrélées au nombre de lits et au nombre de patients ajoutés est vérifiée avec ou sans les points à effet de levier élevé. Cependant, les valeurs prédites et les erreurs-types (E.-T.) diminuent aussi considérablement quand les points à effet de levier élevé sont omis. Par exemple, l'erreur-type sous PPS pour le nombre de lits passe de 13,14 à 6,75 (une réduction de 49 %) ; pour le nombre de patients ajoutés, elle passe de 0,76 à 0,21 (une réduction de 72 %). Ce résultat est dû au fait que certains points ayant un poids extrême sont éliminés dans la régression sous PPS. Par contre, l'erreur-type des estimations par les MCO augmente effectivement quand les points à effet de levier élevé sous les MCO sont omis, parce que la variance d'échantillon des x diminue. Il s'agit d'un autre exemple des écarts importants qui peuvent survenir quand le même type de diagnostic est appliqué aux régressions par les MCO et sous PPS.

Vu que le point 818 est si manifestement extrême, nous avons également ajusté la régression après avoir éliminé cette observation uniquement. Les résultats sont présentés à la section (ii) du tableau 4. Omettre ce point unique cause

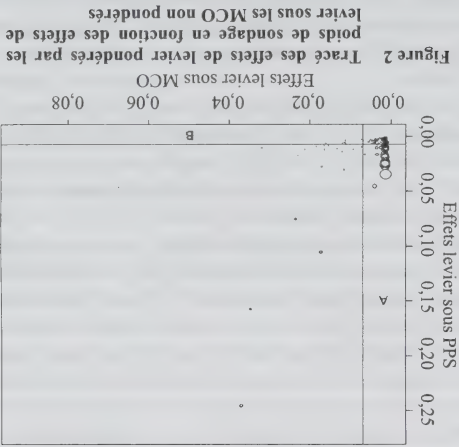
Tableau 2
Observations ayant les 20 effets de levier pondérés les plus importants

ID de l'obs.	h_u sous MCO	h_u pondéré	Poids	Lits	Patients ajoutés
818	0,513	0,389	0,3	49,3	64,7
189	0,037	0,245	3,4	17,7	0,3
346	0,035	0,157	2,2	0,6	16,1
366	0,017	0,105	3,0	0,7	11,1
331	0,024	0,075	1,5	0,1	13,4
271	0,068	0,046	0,4	23,7	0,0
830	0,004	0,045	5,8	5,4	0,1
628	0,056	0,045	0,4	1,0	20,3
179	0,089	0,038	0,2	27,4	0,5
672	0,002	0,034	24,2	1,0	0,8
820	0,048	0,034	0,3	0,8	19,6
207	0,012	0,030	1,3	9,5	0,3
157	0,069	0,030	0,2	23,8	0,5
163	0,017	0,027	0,8	11,4	0,8
613	0,002	0,026	18,5	1,0	0,7
711	0,002	0,024	16,8	1,0	0,9
801	0,002	0,024	17,5	0,6	0,9
156	0,055	0,023	0,2	20,9	0,9
611	0,002	0,023	15,9	1,0	0,8
154	0,051	0,022	0,2	20,5	0,1

Nota: l'ID de l'observation est le numéro de ligne de cette observation dans l'échantillon.

$\bar{w} = 6,57$
 $\bar{x}_1^w = 47,83$
 $\bar{x}_2^w = 1214,13$

Les tailles des poids de sondage peuvent mener les analystes à tirer des conclusions différentes selon qu'ils utilisent les effets de levier pondérés ou non pondérés pour repérer les observations éventuellement influentes. La figure 2 donne un diagramme de dispersion des effets de levier pondérés en fonction de leur version non pondérée. Les deux droites de référence ont été tracées aux valeurs de 0,007. L'observation 818 est omise, parce qu'elle fausserait l'échelle du graphique. Manifestement, les points à effet de levier élevé détectés par la méthode PPS uniquement, situés dans l'aire A, ont tous un poids significativement plus élevés que ceux compris dans l'aire B, qui sont détectés par la méthode de MCO uniquement.



coefficients de pente auraient encore une variance, même si l'on procédait à un recensement. L'esquisse de la justification mathématique de cette façon de faire, qui dépend du modèle (et n'est pas fondée sur le plan) est donnée en annexe

La figure 1 montre les diagrammes de dispersion des dépenses en fonction des nombres de lits et d'ajouts pour l'échantillon de 875 établissements (en omettant un très grand établissement décrit plus loin). Dans la première rangée, les points dont l'effet de levier sous les MCO est supérieur à $2p/n = 0,007$ sont mis en relief. La deuxième rangée donne les graphiques à bulles dans lesquels les bulles sont proportionnelles au poids de chaque cas. Les points à effet de levier élevé sous la pondération par les poids de sondage (PPS) sont mis en relief en utilisant le même seuil de 0,007. Les distributions des variables explicatives sont assez asymétriques comme l'indique le tableau 1. Il existe également un très grand établissement qui n'est pas représenté dans la figure 1 parce qu'il fausse l'échelle du graphique. Pour cet établissement (désigné par observation 818 ici), (dépenses en milliers de dollars; lits; ajouts) = (519 863,3 \$; 2 405; 79 808) et le poids de sondage est égal à 2,22. (L'observation 818 est l'un des cas mentionnés plus haut dont la sélection dans l'échantillon initial était certaine, mais qui ont fait l'objet d'une correction de la non-réponse et qui ont par conséquent un poids final supérieur à 1.) Parce que ses valeurs données sortent fortement de l'alignement de celles obtenues pour les autres organismes, ce point pourrait affecter les estimations.

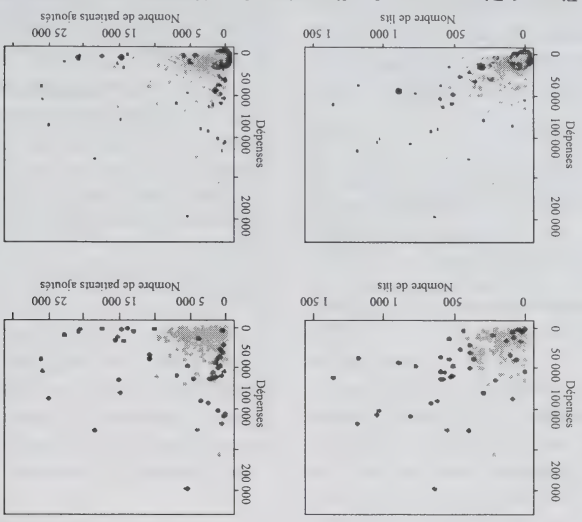


Figure 1 Diagrammes de dispersion des dépenses en fonction du nombre de lits et du nombre de patients ajoutés. Les points dont l'effet de levier est élevé selon les MCO (la PPS) sont mis en relief dans la rangée supérieure (inférieure)

Le tableau 2 donne les 20 observations pour lesquelles les effets de levier sous PPS sont les plus grands. Les valeurs de ces effets de levier varient de 0,022 à 0,389, chiffres considérablement plus élevés que le niveau de 0,007 établi empiriquement. Ce tableau montre aussi, pour ces 20 cas, les effets de levier non pondérés sous les MCO, le ratio de poids de sondage individuel au poids de sondage moyen, ainsi que la distance absolue relative entre les valeurs de X individuelles et leurs moyennes pondérées. Nous notons que l'unité 818 possède les effets de levier pondérés et non pondérés les plus grands, principalement à cause de ses nombres très élevés de lits et de patients ajoutés. Puisque ce cas possède un poids d'échantillonnage inférieur à la moyenne, l'effet de levier sous les MCO est encore plus grand que l'effet pondéré. Il existe d'autres cas semblables, dont les unités 271, 179, 820, 157, 163, 156 et 154, qui sont associées à un nombre extrême de lits, ou un nombre extrême de patients ajoutés, ou aux deux, mais dont les poids sont faibles. Un autre type de valeurs aberrantes découle de poids de sondage extrêmes, même si les valeurs de leurs variables auxiliaires diffèrent peu les unes des autres. Les unités 672, 613, 711, 801 et 611 possèdent toutes un poids de sondage valant plus de 15 fois le poids moyen. Leurs effets de levier pondérés s'avèrent être grands, tandis que les effets de levier non pondérés ne le sont pas. Il existe également un écart appréciable entre les effets de levier pondérés pour le cas 331 ($h_i = 0,075$) et pour le cas 271 ($h_i = 0,046$).

sur toutes les variables du plan de sondage déterminant le schéma d'échantillonnage et qu'il est correct pour la population ainsi que l'échantillon, la régression par les MCO peut être utilisée. Les analystes peuvent formuler des objections à l'intégration des variables du plan de sondage dans un modèle, parce que certaines ne sont pas scientifiquement intéressantes en tant que variables explicatives. En outre, le conditionnement sur toutes les variables du plan de sondage est parfois impossible, surtout quand le « schéma d'échantillonnage » inclut une non-réponse non contrôlée qui, elle-même, peut être liée à la variable réponse. Comme nous le mentionnons dans la section 1, la pondération par les poids de sondage offre un minimum de protection contre la spécification incorrecte du modèle quand la distribution des X dans l'échantillon n'est pas la même que dans la population à cause du type de plan d'échantillonnage utilisé. Néanmoins, certains analystes soutiendront que le plan d'échantillonnage et les poids de sondage peuvent être ignorés dans des applications particulières et que les MCO sont appropriés. Donc, il est intéressant de voir dans quelle mesure les diagnostics pour les MCO et pour la PPS diffèrent dans une application réelle. Cependant, étant donné un mode d'action, un analyste devrait utiliser des diagnostics en harmonie avec la méthode d'ajustement. Si les MCO sont utilisés, les diagnostics standard pour les MCO doivent être examinés; si la régression PPS est utilisée, les diagnostics pour la pondération par les poids de sondage sont ceux qu'il convient d'appliquer. Il se pourrait fort bien que différents points soient influents selon que l'on utilise la régression par les MCO ou la régression PPS.

À la présente section, nous examinons la matrice chapeau et les effets de levier dans un exemple de régression. Nous utilisons la *Survey of Mental Health Organizations (SMHO)* réalisée aux États-Unis en 1998, durant laquelle des données ont été recueillies sur les organismes spécialisés dans les soins de santé mentale et sur les services de soins de santé mentale des hôpitaux généraux. L'échantillon de cette enquête a été tiré selon un plan stratifié à un seul degré avec probabilité proportionnelle à la taille (PPT) (Manderscheid et Henderson 2002; Choudhry 2000). La mesure de taille (MT) utilisée dans l'échantillonnage était le nombre d'épisodes», défini comme étant le nombre de patients/nouveaux patients/sujets ajoutés durant l'année civile 1998. Nombre de variables analysées dans l'enquête sont reliées à la MT et leur distribution non pondérée dans l'échantillon sera différente de la distribution dans la population, puisque l'échantillon a tendance à contenir des unités de plus grande taille. Donc, ce plan est éventuellement informatif, tel qu'il est défini dans Chambers et Skinner (2003).

Les tailles variées des organismes de soins de santé mentale ont produit de grandes fourchettes de valeurs pour les variables étudiées dans l'échantillon, si bien que certaines observations pourraient avoir une influence relativement grande sur les estimations des paramètres d'une régression linéaire. Dans la présente étude, le modèle d'intérêt est la régression des dépenses totales d'un organisme de santé, en millier de dollars, sur le nombre de lits installés et les dépenses ajoutées durant l'année de déclaration. Nous avons utilisé l'estimateur PPS, $\hat{\beta} = (X'W'X)^{-1}X'W'Y$. Imitant la procédure employée par la plupart des analystes, nous n'avons pas intégré de matrice de variance V du modèle dans l'estimation des paramètres de la régression. Au total, nous avons utilisé 875 observations dans la régression, pour chacune desquelles aucune valeur ne manquait pour les variables indépendantes et dépendante.

Le tableau 1 résume les valeurs des quantiles des variables incluses dans la régression, y compris les poids de sondage. Le total des dépenses possède un maximum de 519 863,3, qui vaut près de 30 000 fois le minimum, 16,6. Bien qu'ils ne soient pas aussi extrêmes que le total des dépenses, le nombre de lits et le nombre de patients ajoutés présentent aussi un écart significatif entre leurs valeurs maximale et minimale. Comme l'échantillon a été sélectionné selon un plan PPT, les poids de sondage étaient associés aux tailles des organismes de santé mentale, la fourchette allant de 1 à 158,86. Les poids que nous utilisons dans l'analyse comprennent une correction de la non-réponse qui a été effectuée séparément par strate du plan de sondage. Dans certains cas, les unités qui avaient été sélectionnées avec certitude dans l'échantillon initial n'ont pas répondu et certaines unités sélectionnées avec certitude répondantes avaient un poids corrigé supérieur à 1. En tout, 157 organismes avaient un poids de 1 après la correction de la non-réponse.

Tableau 1
Quantiles des variables dans la régression pour la SMHO

Variables	Quantiles				
	0 %	25 %	50 %	75 %	100 %
Dépenses (milliers de dollars)	16,6	2 932,5	6 240,5	11 842,6	519 863,3
N ^o de lits	0	6,5	36	93	2 405
N ^o d'ajouts	0	558,5	1 410	2 406	79 808
Poids	1	1,42	2,48	7,76	158,86

Dans la régression qui suit, nous avons inclus les unités ayant un poids de 1 dans l'estimation de l'erreur-type. Au lieu de les exclure, conformément à l'approche suivie pour traiter les unités sélectionnées avec certitude dans une estimation purement fondée sur le plan de sondage. Inclure les unités sélectionnées avec certitude concorde avec l'idée qu'un modèle de superpopulation est estimé et que les

Puis, en utilisant le fait que $\hat{N} = m\bar{w}$ avec $\bar{w} = \sum_{i=1}^n w_i / n$, l'effet de levier de la i^e observation, ou le i^e élément diagonal de la matrice chapeau pondérée \mathbf{H} , est

$$h_{ii}^w = \frac{n}{1} \frac{w_i}{w} [1 + N(\mathbf{x}_i - \bar{\mathbf{x}}^w)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}^w)].$$

La forme quadratique, $(\mathbf{x}_i - \bar{\mathbf{x}}^w)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}^w)$, définit un ellipsoïde centré à $\bar{\mathbf{x}}^w$ (par exemple, voir Weisberg 2005, chapitre 8), et $N(\mathbf{x}_i - \bar{\mathbf{x}}^w)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}^w)$ est la distance de Mahalanobis de \mathbf{x}_i à $\bar{\mathbf{x}}^w$. Par conséquent, un effet de levier au poids moyen \bar{w} ou que 2) \mathbf{x}_i est très éloigné de la moyenne pondérée, $\bar{\mathbf{x}}^w$, des \mathbf{X}_i , dans la métrique déterminée par la matrice \mathbf{S} .

Par exemple, dans un modèle linéaire simple ne contenant qu'une seule variable auxiliaire, $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2)$, l'effet de levier de la i^e observation est

$$h_{ii}^w = \frac{1}{w} \frac{w_i}{w} \left[1 + N \frac{\sum_{j=1}^n w_j (x_j - \bar{x}^w)^2}{(x_i - \bar{x}^w)^2} \right].$$

où $\bar{x}^w = \sum_i w_i x_i / N$.

Si les termes d'erreur du modèle ont une structure de variance générale $\varepsilon \sim (0, \mathbf{V})$ et que \mathbf{V} est connu, la matrice chapeau est alors définie par $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}$ avec

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}^T \mathbf{W} \mathbf{V} \mathbf{W}^{-1} \mathbf{I} & \mathbf{I}^T \mathbf{W} \mathbf{V} \mathbf{W}^{-1} \mathbf{X}_1 \\ \mathbf{X}_1^T \mathbf{V}^{-1} \mathbf{W} \mathbf{I} & \mathbf{X}_1^T \mathbf{V}^{-1} \mathbf{W} \mathbf{V} \mathbf{W}^{-1} \mathbf{X}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^s w_i / v_i & \sum_{i=1}^s w_i \mathbf{x}_i^T / v_i \\ \sum_{i=1}^s w_i \mathbf{x}_i / v_i & \sum_{i=1}^s w_i \mathbf{x}_i \mathbf{x}_i^T / v_i \end{pmatrix}.$$

Une formule pour \mathbf{A}^{-1} semblable à celle qui précède s'applique avec $\mathbf{x}^{XX} = \sum_{i=1}^s w_i / N^v$ et $\mathbf{S}^v = \sum_{i=1}^s w_i \mathbf{x}_i \mathbf{x}_i^T / v_i$. Si nous utilisons une \mathbf{V} générale, \mathbf{x}^{XV} et N^v ne sont plus des estimations fondées sur le plan de sondage de \mathbf{T}^{XV} et N , mais sont des estimations de $\mathbf{T}^{XV} = \sum_{i=1}^n \mathbf{x}_i^T / v_i$ et $N^v = \sum_{i=1}^n 1 / v_i$. L'effet de levier de la i^e observation sous ce modèle général est

$$h_{ii}^v = \frac{N^v}{w_i} [1 + N^v (\mathbf{x}_i - \bar{\mathbf{x}}^{wv})^T \mathbf{S}^{v-1} (\mathbf{x}_i - \bar{\mathbf{x}}^{wv})].$$

5. Exemple numérique

Comme nous l'avons mentionné à la section 1, des arguments peuvent être avancés pour justifier le fait d'ignorer les caractéristiques du plan de sondage en général et la pondération en particulier dans l'ajustement des modèles. Grosso modo, quand un modèle est conditionné

Les effets de levier peuvent être décomposés en éléments qui séparent les effets du poids et des valeurs de \mathbf{X} pour une unité. Supposons que le modèle de travail est (1) et qu'il contient une constante, de sorte que

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_T^T \end{pmatrix} \equiv (\mathbf{I} \mathbf{X}_1) \text{ et } \mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_T^T \end{pmatrix},$$

où $\mathbf{x}_i^T = (x_{i1}, \dots, x_{i,p-1})$ représente des vecteurs de dimension $1 \times (p-1)$, \mathbf{I} est un vecteur de dimension $n \times 1$ dont tous les éléments sont égaux à 1, et \mathbf{X}_1 est une matrice de dimensions $n \times (p-1)$. La matrice \mathbf{A} est calculée sous la

forme

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}^T \\ \mathbf{X}_1^T \end{pmatrix} \mathbf{W} (\mathbf{I} \mathbf{X}_1) = \begin{pmatrix} \mathbf{I}^T \mathbf{W} \mathbf{I} & \mathbf{I}^T \mathbf{W} \mathbf{X}_1 \\ \mathbf{X}_1^T \mathbf{W} \mathbf{I} & \mathbf{X}_1^T \mathbf{W} \mathbf{X}_1 \end{pmatrix} \equiv \begin{pmatrix} N & \mathbf{x}^T \\ \mathbf{x} & \mathbf{S} \end{pmatrix} \mathbf{A}_1$$

où \mathbf{x}^T est un vecteur de dimension $(p-1) \times 1$ dont les éléments sont $\mathbf{x}_{ij} = \sum_{i=1}^n w_i x_{ij}$ et \mathbf{A}_1 est une matrice de dimensions $(p-1) \times (p-1)$. En utilisant l'inverse d'une

matrice partitionnée,

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{N}{1} & \frac{N}{1} \mathbf{x}^T \mathbf{S}^{-1} \\ \frac{N}{1} \mathbf{x} \mathbf{S}^{-1} & \frac{N}{1} \mathbf{S}^{-1} - \frac{N}{1} \mathbf{x} \mathbf{S}^{-1} \mathbf{x}^T \mathbf{S}^{-1} \end{pmatrix} = \begin{pmatrix} \frac{N}{1} & \frac{N}{1} \mathbf{x}^T \mathbf{S}^{-1} \\ \frac{N}{1} \mathbf{x} \mathbf{S}^{-1} & \frac{N}{1} \mathbf{S}^{-1} - \frac{N}{1} \mathbf{x} \mathbf{S}^{-1} \mathbf{x}^T \mathbf{S}^{-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \frac{N}{1} \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{x}^T \mathbf{S}^{-1} \\ \mathbf{x} \mathbf{S}^{-1} & \mathbf{S}^{-1} \end{pmatrix} \mathbf{A}^{-1} \mathbf{I}$$

où $\bar{\mathbf{x}}^w = \mathbf{x}^T / N$ est un vecteur de dimension $(p-1) \times 1$, et $\mathbf{S} = \mathbf{A}_1 - \mathbf{x} \mathbf{x}^T / N$ est une matrice de dimensions $(p-1) \times (p-1)$. En simplifiant la matrice chapeau en nous servant de la matrice inverse susmentionnée, nous obtenons

$$\mathbf{H} = \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W}$$

$$= \left\{ \frac{N}{1} \mathbf{I}^T + (\mathbf{X}_1 - \mathbf{I} \bar{\mathbf{x}}^w)^T \mathbf{S}^{-1} (-\bar{\mathbf{x}}^w \mathbf{I}^T + \mathbf{X}_1^T) \right\} \mathbf{W} = \left\{ \frac{N}{1} \mathbf{I}^T + \mathbf{S}^{-1} (\mathbf{x}_1 - \bar{\mathbf{x}}^w, \dots, \mathbf{x}_n - \bar{\mathbf{x}}^w) \right\} \mathbf{W}.$$

$$\beta = (X^T X)^{-1} X^T Y = A^{-1} X^T Y, \quad (2)$$

où $A = X^T X$ est une matrice carrée et inversible. Les valeurs ajustées \hat{Y} correspondant aux valeurs observées Y sont

$$\hat{Y} = X\hat{\beta} = XA^{-1}X^T Y = HY,$$

où $H = XA^{-1}X^T$ est appelé la matrice chapéau. Ce nom a été utilisé pour la première fois par Tukey (Belsley, Kuh et Welsch 1980, chapitre 2 ; Hoaglin et Welsch 1978). L'effet de levier, $h_{ii} = x_i^T A^{-1} x_i$, est le i^e élément sur la diagonale de cette matrice chapéau, qui mesure l'effet de X_i sur sa propre valeur ajustée, puisque $\hat{Y}_i = \sum_j h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$. Si h_{ii} s'approche de 1, X_i joue un rôle crucial dans la détermination de la valeur de Y_i .

La matrice chapéau et les effets de levier basés sur les MCO possèdent de nombreuses propriétés spéciales et utiles :

- i) H est symétrique, ou $h_{ij} = h_{ji}$;
- ii) H est idempotente, ou $H = H^2$, ou

$$(H - H)^2 = 0 ;$$

$$(III) \quad HX = X \text{ ou } (I - H)X = 0 ;$$

$$(IV) \quad 0 \leq h_{ii} \leq 1 ;$$

$$(V) \quad \sum_i h_{ii} = \text{rang}(X) = p, \text{ ce qui implique que l'effet de levier moyen est } \bar{h} = p/n ;$$

si le modèle (I) contient une constante, les deux propriétés suivantes sont vérifiées :

$$(VI) \quad \sum_i h_{ij} = 1 ;$$

$$(VII) \quad h_{ii} = 1/n + (x_i - \bar{x})^T A^{-1} (x_i - \bar{x}), \text{ où } \bar{x} = \sum_i x_i / n.$$

Dans un ensemble de données raisonnablement grand, une valeur d'effet de levier individuel h_{ii} est habituellement considérée comme extrême si elle est au moins égale à deux fois la moyenne, $\bar{h} = p/n$ (Belsley et coll., 1980, chapitre 2). L'existence d'un écart important entre la plupart des cas et quelques cas inhabituels dans la distribution empirique des effets de levier donne aussi la preuve de l'existence d'unités aberrantes.

3. Matrice chapéau pondérée par les poids de sondage

Dans l'approche du pseudo-maximum de vraisemblance, la première étape consiste à former le système d'équations d'estimation qui conviendrait pour un modèle si la population finie entière était observée. Ce système est un type de total de population qui est alors estimé en utilisant des méthodes fondées sur le plan de sondage. Supposons que le modèle structurel sous-jacent est un modèle linéaire à effets fixes :

$$Y_i = x_i^T \beta + \varepsilon_i, \quad \varepsilon_i \sim \text{ind } N(0, \nu_i \sigma^2) \quad (3)$$

où ε_i suit indépendamment une loi normale de moyenne 0 et de variance $\nu_i \sigma^2$, qui est connue à part la constante σ^2 . L'estimateur du pseudo-maximum de vraisemblance (EPMV) de β est la solution du système d'équations d'esti-

maton $X^T W V^{-1} (Y - X\beta) = 0$, avec $V = \text{diag}(\nu_1, \dots, \nu_n)$ et $W = \text{diag}(w_1, \dots, w_n)$. Les poids de sondage, qui, dans les échantillons probabilistes, sont habituellement inversement proportionnels aux probabilités d'inclusion, sont utilisés dans l'EPMV pour tenir compte d'un plan informatif dans lequel la distribution de Y dans l'échantillon diffère vraisemblablement de celle dans la population finie. Ces équations peuvent être résolues explicitement comme $\hat{\beta} = (X^T W V^{-1} X)^{-1} X^T W V^{-1} Y$. Si nous supposons que $V = I$, le modèle (3) se réduit à (1) et l'estimateur pondéré par les poids de sondage (PPS) $\hat{\beta}$ prendra par conséquent la forme d'un estimateur par les moindres carrés pondérés, $\hat{\beta} = (X^T W X)^{-1} X^T W Y$.

S'il est tenu compte des poids de sondage dans la régression, les valeurs prédites deviennent $\hat{Y} = H\hat{Y}$, où la matrice chapéau inclut les poids de sondage et est définie par

$$H = X(X^T W X)^{-1} X^T W = XA^{-1} X^T W$$

avec $A = X^T W X$. Les effets de levier pour la diagonale de la matrice chapéau sont $h_{ii} = x_i^T A^{-1} x_i w_i$. Dans cette formulation, il est supposé que l'analyste n'intègre pas une matrice V dans la régression. Cependant, il est possible de modifier les résultats afin d'intégrer V en utilisant simplement $W' = WV^{-1}$ au lieu de W . Contrairement à la matrice chapéau non pondérée, la matrice chapéau PPS n'est plus symétrique pour les plans d'échantillonnage avec probabilités de sélection inégales (ou, de manière plus générale, babillités de sélection inégales (ii) à (vi) énoncées à la section 2 restent vérifiées (par exemple, voir Valliam, Dorfman et Royall 2000, chapitre 5) à condition que les matrices chapéau non pondérées soient remplacées par leur version pondérée. En outre, la matrice chapéau PPS possède les propriétés supplémentaires utiles, et faciles à vérifier, suivantes :

$$a) \quad WH = WXA^{-1}X^T W = H^T W ;$$

$$b) \quad X^T W(I - H) = X^T W - X^T H^T W = 0 ;$$

$$c) \quad w_i' h_{ii}' = w_i' x_i^T A^{-1} x_i w_i' = w_i' h_{ii}'.$$

La définition des effets de levier pondérés indique qu'un grand effet de levier peut être causé par des valeurs de X aberrantes, un poids aberrant, ou les deux. Notons que les formules de la matrice chapéau et des effets de levier pondérés par les poids de sondage s'appliquent que le plan de sondage soit stratifié, à un degré ou à plusieurs degrés,

modèles ajustés d'après des données d'enquête. Les effets de levier font partie d'une série diagnostique et sont plus efficaces s'ils sont évalués avec les résidus. De nombreuses statistiques diagnostiques, telles que la fameuse distance de Cook (Cook 1977), s'avèrent posséder à la fois des effets de levier et des résidus comme composantes.

Les conseils donnés dans la littérature quant à la façon de traiter les observations influentes une fois qu'elles ont été détectées sont quelque peu ambigus. Une solution évidente, et peut-être naïve, consiste à éliminer les valeurs aberrantes et à rajuster le modèle, ce qui est sensé si ces valeurs aberrantes résultent de données incorrectement enregistrées. Une extension naturelle consisterait à élaborer une approche automatique en vertu de laquelle certaines règles seraient appliquées pour détecter les points influents, les supprimer et réajuster le modèle. Dans le présent article, nous présumons qu'après avoir décelé les points influents et examinés prudemment les raisons de leur influence, un analyste déterminera si les points doivent être exclus de l'ajustement, au lieu d'établir une procédure qui exclurait automatiquement les points sur la base de certaines valeurs seuils.

La suite de l'article est présentée comme il suit. À la section 2, nous décrivons la matrice chapéau obtenue par la méthode des moindres carrés ordinaires, les effets de levier et certaines de leurs propriétés. Aux sections 3 et 4, nous traitons de la matrice chapéau et des effets de levier pondérés par les poids de sondage, ainsi que d'une décomposition qui montre comment les points peuvent posséder un effet de levier important. Les extensions aux données d'enquête s'appliquent aux plans de sondage à un degré ainsi qu'à plusieurs degrés. À la section 5, nous donnons un exemple numérique s'appuyant sur un échantillon à un degré d'organismes spécialisés dans les soins de santé mentale. À la dernière section, nous résumons nos résultats et donnons certaines orientations que pour- raient prendre des travaux de recherche supplémentaires.

2. Matrice chapéau fondée sur les MCO

Un modèle de travail est un modèle qui est considéré provisionnellement par un analyste comme étant la structure qui décrit le mieux une superpopulation conceptuelle. Il peut être révisé après une évaluation plus approfondie par ajout de variables explicatives, suppression de variables explicatives ou d'autres changements ayant trait à sa forme. Supposons que le modèle linéaire de travail est

$$Y = X\beta + \varepsilon, \quad (1)$$
$$Y = (y_1, \dots, y_n)^T, \quad X = (x_1, \dots, x_n)^T, \quad \beta = (\beta_1, \dots, \beta_p)^T, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T.$$

où y_i est la valeur de la variable Y pour l'individu i , x_i est le vecteur de la matrice X pour l'individu i , β est le vecteur des paramètres de la matrice β et ε est le vecteur des résidus. Les moindres carrés ordinaires (MCO) de β est

contre. Les détails peuvent être assez mathématiques et Smith (1989, chapitre 6) exposent les arguments pour et contre. Nous paraphrasons Skinner (2003, section 6.2.3) ici dans (2003, section 6.2.3).

le contexte de l'ajustement d'un modèle linéaire pour prédire une réponse Y basée sur un ensemble de variables explicatives X . Si le modèle linéaire est spécifié correctement et que l'échantillonnage dépend uniquement des variables explicatives comprises dans le modèle, alors les estimations non pondérées des paramètres de régression seront sans biais par rapport au modèle. En particulier, les conditions hypothétiques requièrent que les poids de sondage soient indépendants de X sachant les valeurs des variables explicatives X . Cependant, si l'échantillonnage dépend de facteurs susceptibles d'être reliés à X , même après conditionnement sur les valeurs des variables explicatives, les estimateurs non pondérés des paramètres seront entachés d'un biais à la fois par rapport au modèle vrai et au sens de l'échantillonnage répété, fondé sur le plan de sondage. On dit alors que l'on a un plan de sondage *informatif* dans lequel la distribution des valeurs de X dans l'échantillon diffère de la distribution dans la population. Un exemple de cette situation est donné par Chambers, Dorfman et Sverchkov (2003, section 11.2.3). Si les unités de l'échantillon sont sélectionnées avec des probabilités proportionnelles à une certaine mesure x de leur taille et que X est reliée à x , la distribution de X dans l'échantillon sera étalée vers la droite par rapport à sa distribution dans la population. Cette situation est semblable à celle de l'étude empirique que nous décrivons à la section 5.

L'observation des poids de sondage protégée contre le biais qui pourrait résulter du fait de ne pas tenir compte que l'échantillon est informatif. En outre, si le modèle n'est pas spécifié correctement, la régression pondérée par les poids de sondage estime encore un paramètre de recensement. Autrement dit, les estimations pondérées sont approximativement sans biais pour le modèle linéaire le mieux ajusté qui serait obtenu si l'on disposait de la population finie entière. Dans le présent article, nous supposons qu'un analyste a décidé d'utiliser les poids de sondage dans l'ajustement d'un modèle, éventuellement pour les raisons susmentionnées, et qu'il fournit un type de diagnostic pour évaluer les effets de certains points de données. La matrice chapéau et les effets de levier que nous présentons sont les mêmes que ceux produits par les projections standard quand une régression par les moindres carrés pondérés est exécutée. Cependant, on ne trouve dans la littérature aucune discussion de leur utilisation et de leur interprétation dans le contexte de la régression pondérée par les poids de sondage. Kom et Graubard (1999) est l'une des rares références traitant d'une sorte de diagnostics pour les

Matrice chapéau et effets de levier pondérés par les poids de sondage

Jianzhu Li et Richard Valliant

Résumé

Les diagnostics de régression ont pour objectif de détecter des points individuels ou des groupes de points qui exercent une influence importante sur un modèle ajusté. Lorsqu'on ajuste un modèle à l'aide de données d'enquête, les sources d'influence sont la variable réponse Y , les variables explicatives X et les poids de sondage W . Le présent article traite de l'utilisation de la matrice chapéau et des effets de levier pour détecter les points qui pourraient être influents dans l'ajustement des modèles linéaires parce que les valeurs des variables explicatives ou des poids sont grandes. Nous comparons aussi les résultats qu'un analyste pourrait obtenir s'il utilisait les moindres carrés ordinaires plutôt que les moindres carrés pondérés par les poids de sondage pour déterminer quels points sont influents.

Mots clés : Influence ; régression linéaire ; données d'enquête ; moindres carrés pondérés.

1. Introduction

Dans certains diagnostics de régression linéaire classiques, il est souvent utile de mesurer l'influence que chaque point de donnée peut exercer sur la détermination des valeurs des estimations des paramètres et, à leur tour, des valeurs ajustées. La matrice chapéau et ses éléments diagonaux, appelés effets de levier, sont des méthodes fréquemment utilisées pour repérer les cas possédant des valeurs aberrantes pour les variables explicatives et, par conséquent, susceptibles d'être influents dans l'ajustement du modèle s'ils sont également associés à des résidus inhabituels. Quand la régression comporte plus d'une variable explicative, les analystes peuvent calculer les effets de levier pour résumer l'influence collective des valeurs de X pour chaque observation.

Dans le cas de l'estimation en population finie, les modèles sont habituellement construits en émettant une hypothèse de superpopulation. Supposons qu'un modèle soit raisonnablement bien ajusté pour la majeure partie de la population. Par souci de commodité, nous l'appellerons le modèle « vrai ». Cependant, l'objectif est habituellement de trouver un modèle possédant une certaine puissance descriptive ou prédictive, en se souvenant qu'aucun modèle n'est réellement « vrai ». Les diagnostics d'influence devraient permettre aux analystes de détecter les points qui font s'écarter les paramètres estimés du modèle vrai. Dans la régression linéaire utilisant des données d'enquête complexes, les estimations des paramètres sont souvent calculées par la méthode du pseudo-maximum de vraisemblance, décrite par Skinner, Holt et Smith (1989, chapitre 3) en s'inspirant des idées de Binder (1983). Dans le présent article, nous supposons que l'analyste a décidé qu'un estimateur faisait intervenir les poids de sondage convient pour son problème. Comme nous le montrons plus

loin, la matrice chapéau et les effets de levier pondérés par les poids de sondage sont utiles pour détecter les observations éventuellement influentes causées non seulement par les valeurs de X extrêmes, mais aussi par les grands poids de sondage.

La littérature existante sur les sondages offre des discussions de l'effet des valeurs aberrantes sur certaines estimations d'après des données d'enquête, mais n'accorde que peu d'attention aux diagnostics pour les modèles de régression linéaires. Deville et Särndal (1992) et Porter (1990, 1993) discutent de certains moyens de localiser ou d'étiqueter les poids de sondage extrêmes quand l'objectif est d'estimer des totaux de population et d'autres statistiques descriptives simples. Hülliger (1995) et Moreno-Rebollo, Muñoz-Reyes et Muñoz-Pichardo (1999) abordent la question de l'effet des valeurs aberrantes sur l'estimateur d'Horvitz-Thompson d'un total de population. Smith (1987) fait la démonstration de diagnostics basés sur la suppression de cas et une forme de fonction d'influence. Zaslavsky, Schenker et Belin (2001), ainsi que Beaumont et Alavi (2004) utilisent des stratégies fondées sur l'estimation M pour sous-pondérer les grappes ou les unités influentes. Chambers (1986), Gwet et Rivest (1992), Welsh et Ronchetti (1998), ainsi que Duchesne (1999) étudient des techniques d'estimation robustes aux valeurs aberrantes pour les totaux.

L'une des éternelles questions que se posent les analystes des données d'enquête est celle de savoir s'il faut ou non utiliser les poids de sondage dans l'ajustement des modèles. Les recueils publiés sous la direction de Skinner et coll. (1989) et de Chambers et Skinner (2003) traitent de cette question en détail. Binder et Roberts (2003, chapitre 3), Chambers, Dorfman et Sverchkov (2003, sections 11.2.3, 11.6), Chambers et Skinner (2003, chapitre 1), Korn et Graubard (1999, sections 4.3, 4.4), Pfeffermann (1996) et

- Sndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.
- Sndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Afin d'obtenir une mesure de l'efficacité des trois estimations de variance, nous avons calculé la variance de simulation des 100 000 valeurs de V . Ces variances de simulation sont présentées aux tableaux 10.1, 10.2 et 10.3. Les chiffres sont spectaculairement plus faibles pour $V^{sr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$ que pour les deux autres estimateurs. Le tableau 10.4 montre l'avantage relatif de $V^{sr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$ par rapport à $V^{cr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$. Pour cette population, la variance de simulation de $V^{sr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$ est égale à moins de la moitié de la variance de simulation de $V^{cr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$.

Tableau 10.1

Variance de simulation pour l'estimateur de variance à résidus distincts $V^{sr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$

n_1	3 000	2 000	1 000
4 000	64,82	95,91	484,92
3 000	1 179,62	1 806,79	13 995,94
2 000			

Nota : Les valeurs réelles sont égales aux valeurs présentées multipliées par 10⁶.

Tableau 10.2

Variance de simulation pour l'estimateur de variance à résidu combiné $V^{cr}(Y_{2p\ a\ lin}^{2p\ a\ lin})$

n_1	3 000	2 000	1 000
4 000	153,22	364,08	1 290,41
3 000	2 449,05	6 855,69	33 220,88
2 000			

Nota : Les valeurs réelles sont égales aux valeurs présentées multipliées par 10⁶.

Tableau 10.3

Variance de simulation pour l'estimateur de variance $V^{reg}(Y^{reg})$

n_1	3 000	2 000	1 000
4 000	153,25	364,14	1 289,79
3 000	2 449,36	6 854,52	33 210,31
2 000			

Nota : Les valeurs réelles sont égales aux valeurs présentées multipliées par 10⁶.

Tableau 10.4

Ratio des entrées dans le tableau 10.1 aux entrées correspondantes dans le tableau 10.2

n_1	3 000	2 000	1 000
4 000	0,42	0,26	0,38
3 000	0,48	0,26	0,42
2 000			

Dans une perspective fondée sur le plan de sondage de l'estimation pour les plans d'échantillonnage à deux phases, on peut suivre une approche d'estimation par la régression ou une approche d'estimation par calage. Nous nous concentrons sur l'approche par calage pour créer des estimateurs approximativement sans biais par rapport au plan. La mesure dans laquelle on dispose d'information auxiliaire pour le calage est la clé de l'efficacité des estimations. Nous reconnaissons dans le présent article qu'il existe trois types différents de variables auxiliaires associées aux plans d'échantillonnage à deux phases. Ces trois types ont des caractéristiques d'information différentes. D'après ces caractéristiques, nous définissons quatre vecteurs auxiliaires différents, un pour le calage de première phase et les trois autres pour le calage de deuxième phase. L'approche par calage convient pour analyser les estimateurs résultant de manière systématique. Comme le montre l'article, cette approche donne aussi lieu à un estimateur de variance plus efficace que la méthode classique d'estimation de variance sous des plans d'échantillonnage à deux phases.

11. Discussion

Bibliographie

Axelsson, M. (1998). Variance estimation for the generalised regression estimator under two-phase sampling - a modified approach. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 85-89.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

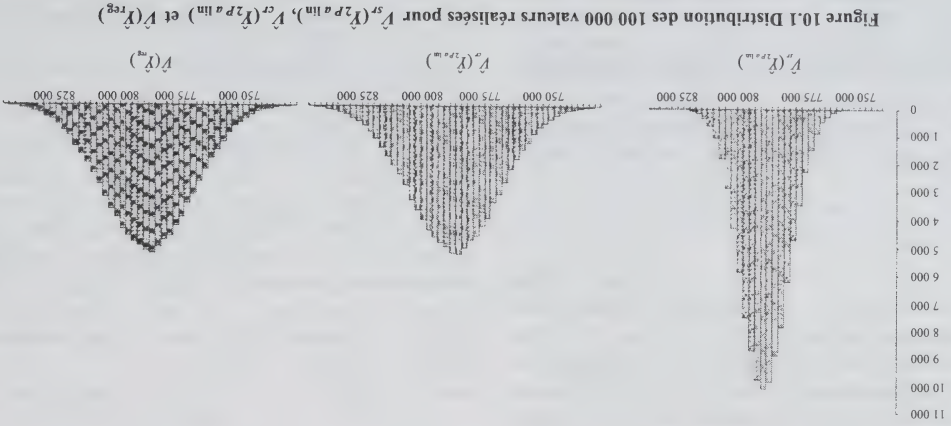
Estevao, V.M., et Samdal, C.-E. (2002). The ten cases of auxiliary information for calibration in two phase sampling. *Journal of Official Statistics*, 18, 233-255.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 157-169.

Hidiroglou, M.A., et Samdal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

Hidiroglou, M.A., Rao, J.N.K. et Haziza, D. (2006). Variance estimation in two phase sampling. (Document accepté à paraître dans) *Australian and New Zealand Journal of Statistics*.

Kott, P.S., et Snijkel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux degrés ? *Techniques d'enquête*, 23, 89-98.

Figure 10.1 Distribution des 100 000 valeurs réalisées pour $V^{sr}(X_{2pa}^{reg})$, $V^{cr}(X_{2pa}^{lin})$ et $V^s(Y^s)$

La figure révèle des distributions qui diffèrent de façon frappante pour $V^{sr}(X_{2pa}^{reg})$ et $V^{cr}(X_{2pa}^{lin})$. La distribution de l'estimateur à résidus distincts $V^{sr}(X_{2pa}^{reg})$ est beaucoup plus concentrée. Donc, $V^{sr}(X_{2pa}^{reg})$ est plus efficace que $V^{cr}(X_{2pa}^{lin})$ et, en moyenne, il produit des intervalles de confiance considérablement plus courts. Nous constatons aussi que la distribution de $V^{cr}(X_{2pa}^{lin})$ est fort semblable à celle de $V^{sr}(X_{2pa}^{reg})$, ce qui appuie l'analyse présentée à la section 9. Nous avons obtenu des résultats semblables pour les autres tailles d'échantillon utilisées dans la simulation.

4 000 et $n = 2 000$.
Pour chaque combinaison (n_1, n_2) , nous avons réalisé 100 000 paires d'échantillons (s_1, s_2) . En nous basant sur les données pour chacun de ces résultats, nous avons calculé l'estimateur de variance à résidus combiné $V^{cr}(X_{2pa}^{lin})$, l'estimateur de variance à résidus distincts $V^{sr}(X_{2pa}^{reg})$ et l'estimateur de variance $V^s(Y^s)$. Pour cela, nous avons utilisé les expressions respectives qui découlent de (7.3), (8.3) et (9.6) quand l'EAS est spécifié à chaque phase. Faute d'espace, les expressions ne sont pas présentées ici. Nous avons obtenu 100 000 valeurs réalisées pour chacun des trois estimateurs de variance. La figure 10.1 montre les distributions pour les 100 000 valeurs de V^s pour $n_1 =$

Nous avons tiré des paires d'échantillons répétés (s_1, s_2) où s_1 est un EAS de n_1 unités provenant de U_1 et s_2 est un

indiqué à la section 9.
faire, nous avons besoin de $\mathbf{x}_k^{(a)} = \phi$, comme il est défini un estimateur $V^{sr}(X_{2pa}^{reg})$ comparable à $V^{cr}(X_{2pa}^{lin})$ et, pour ce pouvons pas comparer $V^{sr}(X_{2pa}^{reg})$ et $V^{cr}(X_{2pa}^{lin})$. Cependant, nous ne comparer $V^{sr}(X_{2pa}^{reg})$ et $V^{cr}(X_{2pa}^{lin})$.
d'avoir $\mathbf{x}_k^{(a)} = \phi$ afin d'exécuter une simulation pour $\sum_{i=1}^n w_i n_{z_k}$. Il importe de noter qu'il n'est pas nécessaire par le vecteur $(N, \sum_{i=1}^n w_i n_{z_k}) = (5 000, 39 611,8)$ $X_{2pa} = \sum_{i=1}^n w_i Y_k$ par calage sur les totaux connus donnés avons déterminé les poids finaux w_i pour l'estimateur Pour chaque échantillon de première phase s_1 , nous

travailler avec X_{2pa}^{reg} et sa forme linéarisée X_{2pa}^{lin} .
 X_{2pa}^{reg} Par conséquent pour cette simulation, nous pouvons conditions pour l'équivalence asymptotique de X_{2pa}^{reg} et $\mathbf{x}_k^{(a)} = \phi$ et $\mathbf{z}_k = \mathbf{x}_k^{(a)}$. Ces spécifications satisfont les $\mathbf{x}_k = (\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \mathbf{x}_k^{(3)})'$ avec $\mathbf{x}_k^{(1)} = (1, u_k)', \mathbf{x}_k^{(2)} = n_{z_k}$. Pour le calage de deuxième phase, nous avons utilisé calage sur le total connu $(N, \sum_{i=1}^n w_i n_{z_k}) = (5 000, 39 611,8)$. dit, les poids w_i pour $k \in s_1$ ont été déterminés par vecteur auxiliaire $\mathbf{x}_k = (1, u_k)'$ et $\mathbf{z}_k = \mathbf{x}_k$. Autrement Pour le calage de première phase, nous avons utilisé le total de population de y donné par $X = \sum_{i=1}^n Y_i = 358 205$. de variance 1. La cible d'estimation dans l'expérience est le ou Normal(0) est la loi normale standard de moyenne 0 et $3n_{z_k} + e_k, k = 1, 2, \dots, 5 000$, avec $e_k \sim 5 \text{Normal}(0)$, les valeurs de la variable d'intérêt selon $Y_k = 10 + n_{z_k} + x_{k-1} - x_k$ pour $x > 0$. En deuxième lieu, nous avons créé indépendantes $u_k \sim 2 \text{Gamma}(4)$ et $n_{z_k} \sim 3 \text{Gamma}(6)$, $2, \dots, 5 000$ par 5 000 réalisations des variables aléatoires nous avons produit les valeurs (u_k, n_{z_k}) pour $k = 1,$

pas à nous inquiéter du choix des poids de départ dans Y_{2p}^{reg} . Nous pouvons simplement travailler avec Y_{2p}^{reg} comme estimateur comparable à Y_{reg}^{reg} et l'estimateur Y_{2p}^{reg} sous ces spécifications.

L'estimateur de variance de Särndal, Swensson et Wretman (1992) contient des facteurs de calage désignés par g_k et g_{1k} . Si nous écartons g_k et g_{1k} , qui ont tous deux une valeur proche de un et dont l'incidence numérique est limitée, leur estimateur de variance est donné par

$$V(Y_{reg}) = \sum_{k \in s} \sum_{l \in s} D_{1k} a_{2k} e_{1ks} e_{1ls} + \sum_{k \in s} \sum_{l \in s} D_{2k} a_{1k} a_{1l} e_{1ks} e_{1ls} \quad (9.6)$$

$$e_{1ks} = y_k - x_{1k}' B_{1s} \text{ et } e_{ks} = y_k - x_k' B_s. \quad (9.7)$$

Les deux composantes de (9.6) sont des doubles sommes

sur s , qui reflètent le fait que e_{1ks} ainsi que e_{ks} ne peuvent être obtenus que pour $k \in s$. La formule (9.6) ressemble à la formule (8.3) pour l'estimateur à résidu combiné, mais à quel point les résidus sont-ils différents dans les deux formules? Examinons les résidus pour l'estimateur ponctuel comparable. Comme nous l'avons mentionné plus haut, cet estimateur Y_{2p}^{reg} possède $x_k' = (x_{1k}', x_{2k}')'$ avec $x_{k(i)} = x_{1k}$ et $x_{k(s)} = x_{2k}$, $\phi_{1k} = \phi_{2k} = x_{1k}' / \sigma_{1k}^2$ et $\phi_{2k} = x_{2k}' / \sigma_{2k}^2$. Sous ces spécifications, les résidus e_{1k} et e_{2k} qui figurent dans (6.1) sont donnés par

$$e_{1k} = x_{2k}' B_{(y;x(2))}^{(y;x(2);x_1)} - x_{1k}' B_{(y;x(1))}^{(y;x(2);x_1)} \text{ pour } k \in s_1$$

$$e_{2k} = y_k - x_k' B_{(y;x)}^{(y;x)}$$

$$= y_k - x_{1k}' B_{(y;x(1))}^{(y;x(2);x_1)} - x_{2k}' B_{(y;x(2))}^{(y;x(2);x_1)} \text{ pour } k \in s \quad (9.8)$$

où $B_{(y;x)}^{(y;x)} = (B_{(y;x(1))}^{(y;x(2);x_1)}, B_{(y;x(2))}^{(y;x(2);x_1)})'$ correspond au partitionnement de $x_k = (x_{1k}', x_{2k}')'$, et de (6.2) il découle que

$$B_{(y;x)}^{(y;x)} = \left(\sum_{s_1} a_k z_k x_{1k}' / \sigma_{1k}^2 \right)^{-1} \left(\sum_{s_1} a_k z_k x_{1k}' x_{2k}' / \sigma_{1k}^2 \right)^{-1} \left(\sum_{s_1} a_k x_{1k}' x_{2k}' / \sigma_{1k}^2 \right)^{-1} \quad (9.9)$$

Les résidus e_{2k} figurant dans (9.8) sont les mêmes que e_{1k} dans (9.7). Mais comment les résidus $e_{12k} = e_{1k} + e_{2k}$ obtenus par addition dans (9.8), sont-ils reliés à leurs homologues e_{1k} dans (9.7)? Pour découvrir ce lien, nous

Pour le démontrer, nous partons de $B_{(y;x)}^{(y;x)}$, qui par définition satisfait $\sum_{s_1} a_k z_k y_k = (\sum_{s_1} a_k z_k x_k') B_{(y;x)}^{(y;x)}$. Cette égalité peut aussi s'écrire sous la forme $\sum_{s_1} a_k z_k y_k = \sum_{s_1} a_k z_k x_{1k}' B_{(y;x(1))}^{(y;x(2);x_1)} + \sum_{s_1} a_k z_k x_{2k}' B_{(y;x(2))}^{(y;x(2);x_1)}$. Puisque $z_k = (x_{1k}' / \sigma_{1k}^2, x_{2k}' / \sigma_{2k}^2)'$, la composante de cette équation correspondant à x_{1k}' / σ_{1k}^2 est $\sum_{s_1} a_k x_{1k}' y_k / \sigma_{1k}^2 = \sum_{s_1} a_k x_{1k}' x_{1k}' B_{(y;x(1))}^{(y;x(2);x_1)} + \sum_{s_1} a_k x_{1k}' x_{2k}' B_{(y;x(2))}^{(y;x(2);x_1)} / \sigma_{1k}^2$. En prémultipliant les deux membres par $(\sum_{s_1} a_k x_{1k}' x_{1k}' / \sigma_{1k}^2)^{-1}$, nous obtenons (9.10).

Puis, en partant de (9.8) et en utilisant la définition de $B_{(y;x(2))}^{(y;x(2);x_1)}$ donnée par (9.9), nous obtenons

$$e_{12k} = e_{1k} + e_{2k}$$

$$= y_k - x_{1k}' B_{(y;x(1))}^{(y;x(2);x_1)} - x_{2k}' B_{(y;x(2))}^{(y;x(2);x_1)}$$

$$= y_k - x_{1k}' \left(B_{(y;x(1))}^{(y;x(2);x_1)} + \left(\sum_{s_1} a_{1k} x_{1k}' x_{1k}' / \sigma_{1k}^2 \right)^{-1} \left(\sum_{s_1} a_{1k} x_{1k}' x_{2k}' B_{(y;x(2))}^{(y;x(2);x_1)} \right) / \sigma_{1k}^2 \right).$$

Dans l'expression entre accolades, remplaçons les deux sommes pondérées par a_k correspondantes sur s ; le résultat est égal à B_{1s} , tel qu'il est donné par (9.10). Cela signifie que $e_{12k} \equiv y_k - x_{1k}' B_{1s} = e_{1ks}$ pour $k \in s$. Donc, l'estimateur de variance (9.6) pour l'estimateur par la régression à deux phases Y_{2p}^{reg} devrait être numériquement proche de l'estimateur de variance à résidu combiné (8.3) pour l'estimateur par le calage Y_{2p}^{reg} défini à la présente section. Nous appuyons empiriquement ceci au moyen de la simulation décrite à la section suivante.

10. Simulation

Dans cette section, nous présentons une petite simulation pour valider l'allégation que l'estimateur de variance à résidus distincts Y_{2p}^{reg} (donné par (7.3)) peut être considérablement plus efficace que l'estimateur de variance à résidu combiné Y_{2p}^{reg} (donné par (8.3)), et que le comportement de ce dernier est semblable à celui de l'estimateur par la régression à deux phases $Y(Y_{reg})$ donné par (9.6). Nous avons créé une population de $N = 5\,000$ unités en deux étapes de la façon suivante : en premier lieu,

9. Comparaison avec l'estimateur par la régression à deux phases

L'estimateur \hat{Y}^{reg} peut-il être interprété comme étant un estimateur par calage ? Pour répondre à cette question, déterminons les poids implicites dans (9.3). Nous pouvons écrire $\hat{Y}^{\text{reg}} = \sum s_k w_k$, avec les poids w_k identifiés en introduisant (9.1) et (9.2) par substitution dans (9.3) et en simplifiant. Nous trouvons que $w_k = a_k g_k$ est donné pour $k \in s$ par

$$g_k = 1 + \left(\sum_{s_1} a_{1k} x_{1k} \right) / \left(\sum_{s_1} a_{1k} x_{1k}^2 / \sigma_1^2 \right) + \left(\sum_{s_1} a_{1k} x_{1k} \right) - \left(\sum_{s_1} a_{1k} x_{1k}^2 / \sigma_1^2 \right) / \left(\sum_{s_1} a_{1k} x_{1k}^2 / \sigma_1^2 \right) \quad (9.4)$$

Les poids w_k ne sont pas énoncés explicitement dans

Särndal, Swensson et Wretman (1992). Dans quel sens, si tant est qu'il y en ait un, pouvons-nous considérer w_k comme un poids de calage ? Pour étudier cette question, nous commençons par remplacer y_k dans (9.3) par x_{1k} . L'utilisation de (9.1) et (9.2) avec $y_k = x_{1k}$ donne $\sum_{s_1} a_{1k} x_{1k}^2$ comme deuxième membre de (9.3). Donc, les poids

$w_k = a_k g_k$ satisfont $\sum_{s_1} w_k x_{1k} = \sum_{s_1} x_{1k}$. Ensuite, nous remplaçons y_k dans (9.3) par x_{2k} , en utilisant de nouveau (9.1) et (9.2) pour obtenir

$$\sum_{s_1} a_{1k} x_{2k} + \left(\sum_{s_1} a_{1k} x_{1k} \right) - \left(\sum_{s_1} a_{1k} x_{1k}^2 / \sigma_1^2 \right) / \left(\sum_{s_1} a_{1k} x_{1k}^2 / \sigma_1^2 \right) \quad (9.5)$$

Bien que (9.5) soit une estimation approximativement

sans biais du total de x_{2k} inconnu $\sum_{s_1} x_{2k}$, elle n'a pas la forme habituelle du deuxième membre d'une équation de calage de deuxième phase, telle que $\sum_{s_1} a_{1k} x_{2k}$ ou $\sum_{s_1} w_{1k} x_{2k}$. Cependant, elle est proche. Si nous remplaçons

les deux sommes sur s_1 par les sommes pondérées appropriées sur s_1 , alors (9.5) devient $\sum_{s_1} w_{1k} x_{2k}$ où w_{1k} est donné par (2.1) avec $z_k = x_{1k} / \sigma_1^2$. Donc, les poids implicites w_k dans \hat{Y}^{reg} donnent le calage exact sur le total de la population de x_{1k} connu et un calage qui est proche de celui sur le total de x_{2k} estimé $\sum_{s_1} w_{1k} x_{2k}$. Cela donne à penser que \hat{Y}^{reg} devrait avoir des propriétés semblables à celles d'un estimateur \hat{Y}^{2p} obtenu en définissant x_k dans

\hat{Y}^{2p} comme étant $x_k = (x_{1k}, x_{2k})'$ avec $x_{k(s)} = x_{1k}$, $x_{k(w)} = x_{2k}$ et $x_{k(o)} = \phi$. En outre, la forme de l'estimateur assisté par modèle implique que $z_k = x_{1k} / \sigma_1^2$ et $z_k = x_k / \sigma_k^2$. Puisque x_k inclut x_{1k} , il est raisonnable de définir $z_k = x_k / \sigma_k^2$ sous la forme $z_k = (x_{1k} / \sigma_1^2, x_{2k} / \sigma_2^2)'$. Ces spécifications satisfont les exigences pour l'équivalence asymptotique de \hat{Y}^{2p} et \hat{Y}^{2p} , de sorte que nous n'avons

9. Comparaison avec l'estimateur par la régression à deux phases

Särndal, Swensson et Wretman (1992) ont élaboré un estimateur par la régression à deux phases pour $Y = \sum y_k$ en se basant sur un article antérieur publié par Särndal et Swensson (1989). Il est utile de voir comment cet estimateur, désigné ici par \hat{Y}^{reg} , se compare à l'estimateur par calage \hat{Y}^{2p} que nous avons considéré dans les sections précédentes. Même s'ils sont fondés sur la même information auxiliaire, les deux estimateurs sont « proches », mais pas identiques, parce que l'estimateur \hat{Y}^{2p} est obtenu par calage à chacune des deux phases d'échantillonnage, tandis que l'estimateur par la régression à deux phases \hat{Y}^{reg} est obtenu par un raisonnement assisté par modèle.

Nous décrivons maintenant l'estimateur par la régression à deux phases de Särndal, Swensson et Wretman (1992). Leur méthode de calcul comprend l'ajustement de deux modèles de régression linéaire à l'aide des données auxiliaires disponibles, l'un au « niveau supérieur » et l'autre au « niveau inférieur ». Ces auteurs ont élaboré un estimateur de variance correspondant, en suivant l'argument de conditionnement classique. Nous comparons leur estimateur de variance avec l'estimateur de variance à résidu combiné (8.3), également obtenu selon l'argument de conditionnement. Les deux estimateurs de variance ne concordent pas exactement, parce que les estimateurs ponctuels sont légèrement différents, mais ils sont numériquement proches,

comme nous le montrons ici. Soit x_{1k} un vecteur de variables auxiliaires dont les totaux de population sont connus et soit $x_k = (x_{1k}, x_{2k})'$, où x_{1k} ainsi que x_{2k} sont des valeurs connues du vecteur pour $k \in s_1$. Nous supposons que le total $\sum_{s_1} x_{1k}$ est connu, tandis que le total $\sum_{s_1} x_{2k}$ est inconnu. Les valeurs prévues produites pour $k \in s_1$ par les deux régressions ajustées au « niveau supérieur » et au « niveau inférieur » sont données respectivement par

$$\hat{y}_{1k} = x_{1k} \hat{b}_{1s} \quad \text{avec} \quad \hat{b}_{1s} = \left(\sum_{s_1} a_{1k} x_{1k} x_{1k} / \sigma_1^2 \right)^{-1} \left(\sum_{s_1} a_{1k} x_{1k} y_k / \sigma_1^2 \right) \quad (9.1)$$

$$\hat{y}_k = x_k \hat{b}_s \quad \text{avec} \quad \hat{b}_s = \left(\sum_{s_1} a_{1k} x_k x_k / \sigma_k^2 \right)^{-1} \sum_{s_1} a_{1k} x_k y_k / \sigma_k^2 \quad (9.2)$$

$$\hat{Y}^{\text{reg}} \text{ de } Y = \sum y_k \text{ est} \quad \hat{Y}^{\text{reg}} = \left(\sum_{s_1} x_{1k} \right) \hat{b}_{1s} + \sum_{s_1^c} a_{1k} (\hat{y}_k - \hat{y}_{1k}) \quad (9.3)$$

Il est simple de définir les estimateurs des deux composantes $V_s(\sum_{i \in s_1} a_i e_{12k})$ et $E_{s_1} V_s(\sum_{i \in s_1} a_i e_{2k})$. Chacune a la forme d'une double somme sur s , parce que e_{12k} et e_{2k} contiennent y_k qui n'est disponible que pour $k \in s$. La première composante utilise e_{12k} pour $k \in s$. Nous obtenons alors $\sum_{k \in s} \sum_{i \in s_1} D_{1k} a_{2k} e_{12k} e_{12i}$ comme estimateur de $V_s(\sum_{i \in s_1} a_i e_{12k})$.

Pour la deuxième composante, nous utilisons les estimations des résidus $\hat{e}_{2k} = y_k - \mathbf{x}_k^k \mathbf{B}^{(y;x)}$ données par (6.1) pour $k \in s$, et nous obtenons $\sum_{k \in s} \sum_{i \in s_1} D_{2k} a_{1k} a_{1i} e_{2k} \hat{e}_{2i}$ comme estimateur de $E_{s_1} V_s(\sum_{i \in s_1} a_i e_{2k})$. Par sommation des deux termes estimés, nous obtenons l'estimateur de variance qui suit, où l'indice inférieur cv pour (*combined residual*) signifie « résidu combiné ».

$$\hat{V}^{cv}(\hat{Y}^{2pa \text{ lin}}) = \sum_{k \in s} \sum_{i \in s_1} D_{1k} a_{2k} e_{12k} e_{12i} + \sum_{k \in s} \sum_{i \in s_1} D_{2k} a_{1k} a_{1i} e_{2k} \hat{e}_{2i}. \quad (8.3)$$

Examinons maintenant en quoi (7.3) et (8.3) diffèrent.

L'estimateur de variance à résidus distincts (7.3) a pour point de départ le développement $V(\sum_{i \in s_1} a_i e_{1k}) = V(\sum_{i \in s_1} a_i e_{1k}) + 2 \text{Cov}(\sum_{i \in s_1} a_i e_{1k}, \sum_{i \in s_1} a_i e_{2k})$. Nous estimons ces trois composantes séparément en tant que fonctions des résidus e_{1k} et e_{2k} . L'expression résultante de la variance contient trois termes : une double somme sur s_1 en ce qui a trait à e_{1k} et e_{1i} , une double somme sur s en ce qui a trait à e_{2k} et e_{2i} , ainsi qu'une somme croisée sur s_1 et s en ce qui a trait à $e_{1k} \in s_1$ et $e_{2i} \in s$. Enfin, nous arrivons à (7.3) en estimant e_{1k} par \hat{e}_{1k} pour $k \in s_1$ et e_{2k} par \hat{e}_{2k} pour $k \in s$.

L'estimateur de variance à résidu combiné (8.3) découle du conditionnement classique sur l'échantillon de première phase s_1 , sous la forme $V(\hat{Y}^{2pa \text{ lin}}) = V(E_{s_1} \hat{Y}^{2pa \text{ lin}}) + E_{s_1} V(\hat{Y}^{2pa \text{ lin}})$. Cela nous amène à combiner e_{1k} et e_{2k} sous la forme $e_{12k} = e_{1k} + e_{2k}$ dans le premier terme. Le deuxième terme, $E_{s_1} V(\hat{Y}^{2pa \text{ lin}})$, est une fonction de e_{2k} . Puisque e_{12k} et e_{2k} ne peuvent être estimés que sur s , l'estimateur de variance résultant devient une somme de deux termes, exprimés chacun sous la forme d'une double somme sur s .

L'estimateur à résidus distincts (7.3) est plus efficace que l'estimateur à résidu combiné (8.3), parce qu'il est fondé sur plus grand s_1 . L'avantage de (7.3) sur (8.3) est illustré par la simulation décrite à la section 10. L'approche qui sous-tend l'estimateur de variance à résidus distincts (7.3) peut être étendue à l'échantillonnage à deux phases et à d'autres plans de sondage complexes. Dans ces extensions de la méthode, nous procédons de manière semblable, en commençant par l'établissement de la forme linéaire par développement des composantes de variance et la détermination des résidus appropriés.

$$\hat{V}^{sr}(\hat{Y}^{2pa \text{ lin}}) = \sum_{k \in s_1} \sum_{i \in s_1} D_{1k} a_{1k} e_{1k} e_{1i} + \sum_{k \in s} \sum_{i \in s_1} D_{2k} a_{2k} e_{2k} \hat{e}_{2i} + 2 \sum_{k \in s_1} \sum_{i \in s_1} D_{1k} a_{2k} e_{1k} \hat{e}_{2i}. \quad (7.3)$$

Le terme « résidus distincts » et l'indice inférieur correspondant sr pour (*separate residual*) reflètent le fait que, dans (7.3), les résidus demeurent distincts, \hat{e}_{1k} étant défini sur le grand échantillon s_1 et \hat{e}_{2k} sur l'échantillon plus petit s . Le fait que les résultats calculés pour le grand échantillon s_1 peuvent être avantageux pour l'estimation de la variance de calcul diffère de notre approche par calage fondée sur \mathbf{x}_k et \mathbf{x}_k^k . La méthode d'estimation de la variance de l'estimateur par la régression à deux phases présentée dans Hidroglou, Rao et Haziza (2006) possède certains traits en commun avec notre approche, mais il existe aussi d'importantes différences.

8. Estimateur de la variance à résidu combiné

Nous arrivons à l'expression (7.3) en reconnaissant que les estimations \hat{e}_{1k} peuvent être obtenues pour $k \in s_1$. L'approche classique, passée en revue ici, consiste à obtenir un estimateur de variance par conditionnement sur l'échantillon de première phase s_1 . On obtient un estimateur de variance où tous les résidus requis sont définis pour $k \in s$. Plus tard, nous le comparerons à l'estimateur plus efficace (7.3). Partant de (5.1), nous conditionnons sur l'échantillon de première phase s_1 pour obtenir

$$V(\hat{Y}^{2pa \text{ lin}}) = V(E_{s_1} \hat{Y}^{2pa \text{ lin}}) + \sum_{k \in s} a_k e_{2k} (E_{s_1} V(\hat{Y}^{2pa \text{ lin}}) + \sum_{k \in s} a_k e_{1k} + \sum_{k \in s} a_k e_{2k})$$

$$e_{12k} = y_k - \mathbf{x}_k^{k(i)} \mathbf{B}^{(y;x(i))} - \mathbf{x}_k^{k(w)} \mathbf{B}^{(y;x(w))}$$

où $e_{12k} = e_{1k} + e_{2k}$ est appelé le résidu combiné. À partir de (4.10), nous obtenons les expressions suivantes.

$$e_{12k} = y_k - \mathbf{x}_k^{k(i)} \mathbf{B}^{(y;x(i))} - \mathbf{x}_k^{k(w)} \mathbf{B}^{(y;x(w))} - \mathbf{x}_k^{k(a)} \mathbf{B}^{(y;x(a))}. \quad (8.2)$$

Ici, $\pi_{2k\ell}$ et $a_{2k\ell}$ sont conditionnés sur l'échantillon s_1 . Nous supposons que toutes les probabilités d'inclusion de premier et de deuxième ordre sont positives. En utilisant cette notation et les résultats susmentionnés, nous allons maintenant développer deux estimateurs de variance différents aux deux sections suivantes.

7. Estimateur de variance avec résidus distincts

La variance de $Y_{2pa\text{lin}}$ est donnée par (5.1), où e_{1k} et e_{2k} sont définis par (4.10). Elle peut être développée sous la forme

$$V(Y_{2pa\text{lin}}) = V\left(\sum_{s_1} a_{1k} e_{1k}\right) + V\left(\sum_{s_1} a_{2k} e_{2k}\right) + 2 \text{Cov}\left(\sum_{s_1} a_{1k} e_{1k}, \sum_{s_1} a_{2k} e_{2k}\right). \quad (7.1)$$

Si nous connaissons les résidus e_{1k} et e_{2k} , des estimations sans biais de ces trois composantes seraient données, respectivement, par

$$\sum_{k \in s_1} D_{1k\ell} e_{1k} e_{1\ell}, \quad \sum_{k \in s_2} D_{2k\ell} e_{2k} e_{2\ell}, \quad 2 \sum_{k \in s_1} \sum_{\ell \in s_2} D_{1k\ell} a_{2\ell} e_{1k} e_{2\ell}. \quad (7.2)$$

La preuve de l'absence de biais est semblable pour les trois composantes. Par exemple, pour la deuxième, nous

$$E_{s_1} E_{s_2 | s_1} \left(\sum_{k \in s_2} \sum_{\ell \in s_2} D_{2k\ell} e_{2k} e_{2\ell} \right)$$

$$= E_{s_1} \left(\sum_{k \in s_1} \sum_{\ell \in s_1} (D_{1k\ell} / a_{2k\ell}) e_{2k} e_{2\ell} \right)$$

$$= \sum_{k \in U} \sum_{\ell \in U} (D_{1k\ell} / a_{2k\ell}) e_{2k} e_{2\ell}$$

$$= \sum_{k \in U} \sum_{\ell \in U} (a_{2k\ell} / a_{1k\ell}) e_{2k} e_{2\ell} - \left(\sum_{\ell \in U} e_{2\ell} \right)^2$$

$$= E \left[\left(\sum_{s_1} a_{2k\ell} \right)^2 \right] - E \left[\left(\sum_{s_1} a_{2k\ell} \right)^2 \right]$$

$$= V \left(\sum_{s_1} a_{2k\ell} \right).$$

Nous remplaçons maintenant dans (7.2) les résidus inconnus par les estimations respectives données par (6.1), c'est-à-dire e_{1k} par \hat{e}_{1k} pour $k \in s_1$ et e_{2k} par \hat{e}_{2k} pour $k \in s_2$. Puis, nous additionnons les trois composantes résultantes pour obtenir l'estimateur de variance avec « résidus distincts »

$$\begin{aligned} \hat{e}_{1k} &= \mathbf{x}_k' \mathbf{B}^{(y;x(a))} + \mathbf{x}_k' \mathbf{B}^{(y;x(w))} \quad \text{for } k \in s_1 \\ &\quad - \mathbf{x}_k' \mathbf{B}^{(x(b(w);x_1))} \quad \text{for } k \in s_1 \\ \hat{e}_{2k} &= y_k - \mathbf{x}_k' \mathbf{B}^{(y;x)} \\ &\quad = y_k - \mathbf{x}_k' \mathbf{B}^{(y;x(\ell))} - \mathbf{x}_k' \mathbf{B}^{(y;x(w))} \quad \text{for } k \in s \end{aligned} \quad (6.1)$$

$$\text{et} \quad \mathbf{B}^{(x(b(w);x_1))} = \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}' \right)^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}' \mathbf{B}^{(y;x(w))}. \quad (6.2)$$

Dans la définition de \hat{e}_{1k} , le terme $\mathbf{B}^{(x(b(w);x_1))}$ est l'esti-

mation de $\mathbf{B}^{(x(b(w);x_1))} = \left(\sum_{s_1} \mathbf{z}_{1k} \mathbf{x}_{1k}' \right)^{-1} \sum_{s_1} \mathbf{z}_{1k} \mathbf{x}_{1k}' \mathbf{B}^{(y;x(w))}$ dans (4.10). Deux remplacements sont requis dans $\mathbf{B}^{(x(b(w);x_1))}$ pour arriver à $\mathbf{B}^{(x(b(w);x_1))}$: pour commencer, les sommes sur U sont remplacées par les sommes pondérées appropriées sur s_1 , ce qui donne $\mathbf{B}^{(x(b(w);x_1))} = \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}' \right)^{-1} \sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}' \mathbf{B}^{(y;x(w))}$. Dans cette expression, $\mathbf{B}^{(y;x(w))}$ est encore inconnu, de sorte que nous le rempla-

çons par son estimation $\mathbf{B}^{(y;x(w))}$ pour arriver à $\mathbf{B}^{(x(b(w);x_1))}$.

Un point important à souligner est que les estimations \hat{e}_{1k} peuvent être obtenues pour $k \in s_1$, parce que \mathbf{x}_{1k} et $\mathbf{x}_{k(w)}$ sont toutes connues pour $k \in s_1$, mais que les estimations \hat{e}_{2k} ne peuvent être calculées que pour $k \in s_2$, parce que y_k n'est disponible que pour $k \in s_2$. Le fait que les estimations \hat{e}_{1k} soient disponibles pour $k \in s_1$ plutôt que pour $k \in s_2$ nous permet de construire (à la section 7) un estimateur plus efficace de $V(Y_{2pa\text{lin}})$ que celui obtenu selon la méthode classique d'estimation de la variance (à la section 8) où tous les résidus estimés sont calculés uniquement pour $k \in s_2$.

Les poids de sondage $a_{1k} = 1/\pi_{1k}$, $a_{2k} = 1/\pi_{2k}$ et $a_{k\ell} = a_{1k} a_{2\ell}$ sont définis à la section 1. Dans les sections qui suivent, nous avons également besoin des quantités données ci-dessous, définies comme les fonctions des probabilités d'inclusion de deuxième ordre $\pi_{1k\ell} = \Pr(k \& \ell \in s_1) = \Pr(k \& \ell \in s_2 | s_1)$:

$$\begin{aligned} a_{k\ell} &= 1/\pi_{1k\ell}, \quad a_{2k\ell} = 1/\pi_{2k\ell}, \quad a_{k\ell} = a_{1k\ell} a_{2\ell\ell} \\ D_{1k\ell} &= a_{1k} a_{1\ell} - a_{1k\ell}, \quad D_{2k\ell} = a_{2k} a_{2\ell} - a_{2k\ell}, \\ D_{k\ell} &= a_{k\ell} a_{\ell\ell} - a_{k\ell}. \end{aligned}$$

comparativement à ceux obtenus en exécutant le calage de deuxième phase. La forme linéarisée de l'estimateur à deux phases avec w_k s'obtient en l'écrivant sous la

forme

$$\hat{Y}_{2p} = \mathbf{X}_1' \mathbf{B}_{(Y; \mathbf{x}_1)} - \hat{\mathbf{X}}_1' \mathbf{B}_{(Y; \mathbf{x}_1)} + \sum_s a_k Y_k \\ + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(Y; \mathbf{x}_1) s_1} - \mathbf{B}_{(Y; \mathbf{x}_1)}) \\ + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \left(\sum_{s_1} a_k \mathbf{z}_{1k} \mathbf{x}_1' \right)^{-1} \\ \left(\sum_s a_k \mathbf{z}_{1k} Y_k - \sum_{s_1} a_k \mathbf{z}_{1k} Y_k \right).$$

Les termes $\mathbf{B}_{(Y; \mathbf{x}_1)}$ et $\mathbf{B}_{(Y; \mathbf{x}_1) s_1}$ sont définis à la section précédente. Quand les échantillons sont suffisamment grands, $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\sum_{s_1} a_k \mathbf{z}_{1k} \mathbf{x}_1')^{-1} (\sum_s a_k \mathbf{z}_{1k} Y_k - \sum_{s_1} a_k \mathbf{z}_{1k} Y_k)$ et $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(Y; \mathbf{x}_1) s_1} - \mathbf{B}_{(Y; \mathbf{x}_1)})$ sont d'ordre plus faible et peuvent être ignorés. Nous obtenons ainsi la forme linéarisée de l'estimateur

$$\hat{Y}_{2p \text{ lin}} = \mathbf{X}_1' \mathbf{B}_{(Y; \mathbf{x}_1)} - \hat{\mathbf{X}}_1' \mathbf{B}_{(Y; \mathbf{x}_1)} + \sum_s a_k Y_k. \quad (4.13)$$

Nous pouvons également écrire cette forme linéarisée comme une somme de trois termes résiduels, les résidus e_{0k} , e_{1k} et e_{2k} ayant les définitions suivantes pour

$$e_{0k} = \mathbf{x}_{1k}' \mathbf{B}_{(Y; \mathbf{x}_1)} \\ e_{1k} = -\mathbf{x}_{1k}' \mathbf{B}_{(Y; \mathbf{x}_1) s_1}$$

$$e_{2k} = Y_k.$$

Ces résidus ressemblent à ceux donnés par (4.10) si nous posons que $\mathbf{x}_k = \phi$ et que nous supprimeons $\mathbf{B}_{(Y; \mathbf{x}_1)}$. Notons que $\mathbf{B}_{(Y; \mathbf{x}_1)}$ joue le même rôle que $\mathbf{B}_{(\mathbf{x}^{(w)}; \mathbf{x}_1)}$ dans (4.10). Comme auparavant, $e_{0k} + e_{1k} + e_{2k} = Y_k$ pour chaque k , d'où $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U Y_k = Y$. L'estimateur à double facteur d'extension est un cas particulier de cet estimateur quand nous avons aussi $\mathbf{x}_{1k} = \phi$. Cela signifie que $\mathbf{B}_{(Y; \mathbf{x}_1)}$ n'est pas défini. Les définitions correspondantes pour e_{0k} , e_{1k} et e_{2k} sont simplement $e_{0k} = 0$, $e_{1k} = 0$ et $e_{2k} = Y_k$ pour $k \in U$.

Aux sections qui suivent, nous examinons le biais et la variance de l'estimateur par calage à deux phases \hat{Y}_{2p} et nous proposons une nouvelle méthode d'estimation de la variance de ces deux groupes d'estimateurs ont des propriétés et une forme linéarisée semblables. La seule différence tient à l'estimation de la variance. Nous utilisons le même estimateur de variance (décrit à la section 7), et nous

5. Biais et variance de l'estimateur par calage à deux phases \hat{Y}_{2p}

L'estimateur par calage à deux phases $\hat{Y}_{2p} = \sum_s w_k Y_k$ est approximativement sans biais pour $Y = \sum_U Y_k$. Pour le montrer, nous calculons l'espérance de la forme linéarisée donnée par (4.9) par la méthode habituelle de conditionnement sur l'échantillon de première phase s_1 . Nous avons

$$E(\sum_s a_k e_{2k}) = E(s_1' E_{s_1 | s_1} (\sum_s a_k e_{2k})) = \sum_U e_{0k}.$$

Cela montre que $\hat{Y}_{2p \text{ lin}}$ est sans biais pour Y . En vertu de (4.4), $\hat{Y}_{2p} = \hat{Y}_{2p \text{ lin}} + R$, de sorte que le biais de \hat{Y}_{2p} est égal à l'espérance de R , qui est la somme des deux termes d'ordre inférieur $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\mathbf{B}_{(\mathbf{x}^{(w)}; \mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}^{(w)}; \mathbf{x}_1) s_1})$ et $(\mathbf{X} - \hat{\mathbf{X}})' (\mathbf{B}_{(Y; \mathbf{x}_1)} - \mathbf{B}_{(Y; \mathbf{x}_1) s_1})$. Comme nous l'avons mentionné à la section 4, l'espérance de chacun de ces termes est presque nulle. Il s'ensuit que \hat{Y}_{2p} est approximativement sans biais pour Y .

La variance de $\hat{Y}_{2p} = \sum_s w_k Y_k$ est approximée de près par la variance de la forme linéarisée $\hat{Y}_{2p \text{ lin}}$ donnée par (4.9) avec les résidus définis par (4.10). Son premier terme, $\sum_U e_{0k}$, est constant et ne contribue pas à la variance. Donc,

$$V(\hat{Y}_{2p \text{ lin}}) = V\left(\sum_{s_1} a_k e_{1k} + \sum_s a_k e_{2k}\right). \quad (5.1)$$

Nous utilisons (5.1) comme point de départ pour déterminer un estimateur de variance pour $\hat{Y}_{2p \text{ lin}}$. Deux approches distinctes peuvent être suivies et leur comparaison est intéressante. Celle décrite à la section 7 est nouvelle et plus intéressante que celle décrite à la section 8, obtenue par la méthode classique de conditionnement sur l'échantillon de première phase s_1 , parce qu'elle produit un estimateur de variance plus efficace. Les résidus e_{1k} et e_{2k} donnés par (4.10) jouent un rôle important dans les deux dérivations.

6. Préliminaires de l'estimation de la variance

Notre objectif est d'estimer la variance $V(\hat{Y}_{2p \text{ lin}})$ donnée par (5.1). Nous le faisons aux sections 7 et 8 en nous servant de deux arguments différents. Les résidus e_{1k} et e_{2k} sont définis pour tout $k \in U$, mais ils ne peuvent pas être calculés. Ils doivent être remplacés par les estimations \hat{e}_{1k} et \hat{e}_{2k} . Ces estimations, formées à l'image de (4.10), sont

Si nous comparons (4.5) et (4.8), nous constatons que $Y_{2p\text{ lin}}^{2p\text{ lin}} = Y_{2p\text{ lin}}^{2p\text{ lin}} + (X_1 - \bar{X}_1)' B_{(Y;X)} - B_{(X;X)}' B_{(Y;X)}$, comme il est énoncé dans le résultat. Cela complète la preuve du résultat 4.1.

Le résultat 4.1 montre qu'en général, les formes linéaires de $Y_{2p\text{ w}}^{2p\text{ w}}$ et $Y_{2p\text{ a}}^{2p\text{ a}}$ ne sont pas les mêmes. Cependant, elles le sont sous certaines conditions. Considérons le cas du calage emboîté (à ne pas confondre avec l'échantillonnage emboîté), ce qui signifie que x_k inclut x_{1k} . Alors, x_k est de la forme $x_k = (x_{1k}, x_{+k})'$, où le vecteur x_{+k} est composé des variables restantes. Nous énonçons et prouvons maintenant le résultat suivant.

Résultat 4.2 : Si $x_k = (x_{1k}, x_{+k})'$ et $z_k = (z_{1k}, z_{+k})'$, alors $Y_{2p\text{ w lin}}^{2p\text{ w lin}} = Y_{2p\text{ a lin}}^{2p\text{ a lin}}$ et $Y_{2p\text{ w}}^{2p\text{ w}} = Y_{2p\text{ a}}^{2p\text{ a}}$ sont asymptotiquement équivalents.

Preuve

La preuve découle du résultat 4.1 en montrant que $B_{(Y;X)} - B_{(X;X)}' B_{(Y;X)} = 0$ sous les conditions spécifiées.

Nous avons

$$B_{(Y;X)} - B_{(X;X)}' B_{(Y;X)} = \left(\sum U z_{1k} x_{1k}' \right)^{-1} \left(\sum U z_{1k} h_k \right)$$

où $h_k = y_k - x_k' (\sum U z_k x_k')^{-1} (\sum U z_k y_k)$. Puisque $\sum U z_{1k} h_k = 0$ et que nous nous disposons que $z_k = (z_{1k}, z_{+k})'$, il s'ensuit que $\sum U z_{1k} h_k = 0$ et $B_{(Y;X)} - B_{(X;X)}' B_{(Y;X)} = 0$. Donc, il découle du résultat 4.1 que $Y_{2p\text{ w lin}}^{2p\text{ w lin}} = Y_{2p\text{ a lin}}^{2p\text{ a lin}}$. Puisque leurs formes linéaires sont les mêmes, $Y_{2p\text{ w}}^{2p\text{ w}}$ et $Y_{2p\text{ a}}^{2p\text{ a}}$ sont des estimateurs asymptotiquement équivalents.

Il se trouve que le résultat 4.2 requiert seulement que nous incluions x_{1k} dans la composante $x_k^{(i)}$. De toute évidence, il est logique d'inclure x_{1k} dans la composante $x_k^{(i)}$ de x_k , parce que les totaux de x_1 sont connus. Cependant, nous obtenons le même résultat asymptotique à condition que toutes les variables comprises dans x_{1k} soient incluses quelque part dans $x_k = (x_{1k}^{(i)}, x_{+k}^{(w)})' = x_k^{(i)}$. En pratique, nous constatons souvent que $x_k^{(i)} = x_{1k}$ avec $z_{1k} = x_{1k}$ et $z_k = x_k = (x_{1k}, x_{+k})'$ où x_{+k} est le vecteur des variables restantes $x_k^{(w)}$ et $x_k^{(a)}$. Cela satisfait les exigences pour l'équivalence asymptotique de $Y_{2p\text{ a}}^{2p\text{ a}}$ et $Y_{2p\text{ w}}^{2p\text{ w}}$. Pour étudier les propriétés de $Y_{2p\text{ a}}^{2p\text{ a}}$ et $Y_{2p\text{ w}}^{2p\text{ w}}$, nous nous servons des formes linéaires données respectivement par (4.5) et (4.8). Moyennant des définitions appropriées pour les résidus e_{0k}, e_{1k} et e_{2k} , nous pouvons représenter $Y_{2p\text{ a lin}}^{2p\text{ a lin}}$ et $Y_{2p\text{ w lin}}^{2p\text{ w lin}}$ comme la somme de trois termes : un terme constant $\sum U e_{0k}$, un terme d'extension de première phase $\sum s_{1k} e_{1k}$ et un terme d'extension double $\sum s_{2k} e_{2k}$.

$$Y_{2p\text{ lin}}^{2p\text{ lin}} = \sum U e_{0k} + \sum s_{1k} e_{1k} + \sum s_{2k} e_{2k}. \quad (4.9)$$

Cela fait de (4.9) un point de départ convenable pour étudier le biais et la variance asymptotique des deux estimateurs $Y_{2p\text{ a lin}}^{2p\text{ a lin}}$ et $Y_{2p\text{ w}}^{2p\text{ w}}$. Pour la forme linéaire $Y_{2p\text{ a lin}}^{2p\text{ a lin}}$ donnée par (4.5), les trois quantités résiduelles sont définies comme il suit pour $k \in U$:

$$e_{0k} = x_k^{(i)} B_{(Y;X)} + x_{1k}' B_{(X;X)} - x_k^{(a)} B_{(Y;X)}$$

$$e_{1k} = x_k^{(a)} B_{(Y;X)} + x_{1k}' B_{(X;X)} - x_k^{(w)} B_{(Y;X)}$$

$$- x_{1k}' B_{(X;X;X_1;X_1)}$$

$$e_{2k} = y_k - x_k^{(i)} B_{(Y;X)} - x_{1k}' B_{(X;X)} - x_k^{(w)} B_{(Y;X)}$$

$$- x_{1k}' B_{(Y;X)} - x_k^{(a)} B_{(Y;X)}$$

Nous que e_{2k} est simplement $e_{(Y;X)}$. De même, pour $Y_{2p\text{ w lin}}^{2p\text{ w lin}}$ donné par (4.8), les résidus ont les définitions suivantes pour $k \in U$:

$$e_{0k} = x_k^{(i)} B_{(Y;X)}$$

$$+ x_{1k}' B_{(X;X)} + B_{(Y;X)} - B_{(X;X)}' B_{(Y;X)}$$

$$e_{1k} = x_k^{(a)} B_{(Y;X)} + x_{1k}' B_{(X;X)} - x_k^{(w)} B_{(Y;X)}$$

$$- x_{1k}' B_{(X;X)} + B_{(Y;X)} - B_{(X;X)}' B_{(Y;X)}$$

$$e_{2k} = y_k - x_k^{(i)} B_{(Y;X)}$$

Notons que, dans les deux cas, $e_{0k} + e_{1k} + e_{2k} = y_k$ pour chaque k , d'où $\sum U (e_{0k} + e_{1k} + e_{2k}) = \sum U y_k = Y$. Cette additivité nous permet de prouver à la section 5 que $Y_{2p\text{ a}}^{2p\text{ a}}$ et $Y_{2p\text{ w}}^{2p\text{ w}}$ sont approximativement sans biais. Faute d'espace, dans les sections qui suivent, nous nous concentrons sur les propriétés de $Y_{2p\text{ a}}^{2p\text{ a}}$. Cependant, l'analyse est la même pour $Y_{2p\text{ w}}^{2p\text{ w}}$ et la méthode d'estimation de la variance proposée à la section 7 peut également être utilisée pour cet estimateur.

4.2 Estimateurs sans calage de deuxième phase

$$(x_k = \phi)$$

En l'absence de calage de deuxième phase ($x_k = \phi$), nous avons $w_k = a_k^*$. Conséquemment, les poids finaux sont soit $w_k = a_k^* a_{2k}$ soit $w_k = w_{1k} a_{2k}$. La première option donne l'estimateur à double facteur d'extension $\sum s_{2k} a_k y_k$. La deuxième produit un estimateur différent qui est habituellement moins efficace. Cependant, ces estimateurs sont généralement tous deux inefficaces

$$Y_{2p,a}^{2p,w} = \sum^U (X_i^{(a)} B^{(a)}(x^{(a)}) + x_i' B^{(a)}(x^{(a)})) + \sum^{s_1} a_{1k}^{(a)} (X_i^{(a)} B^{(a)}(x^{(a)}) + e^{(a)}(x^{(a)}; x_1)) + (X_1 - X_1') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) + (X - X') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) \quad (4.4)$$

Le terme $B^{(a)}(x^{(a)}) = \sum^U a_k z_k x_k^{(a)} - \sum^S a_k z_k y_k$ converge en probabilité vers (et est approximativement sans biais pour) $B^{(a)}(x^{(a)}) = (\sum^U z_k x_k^{(a)})^{-1} \sum^U z_k x_k^{(a)}$. Le premier terme est constant et ne contribue pas à la variance de $Y_{2p,a}^{2p,w}$. Les

deux prochains termes du milieu sont des quantités aléatoires, définies comme étant des sommes sur s_1 et s_2 , respectivement. Les deux derniers termes sont des produits de différences dont l'espérance est nulle ou presque

nulle. Pour ce qui est du produit $(X_1 - X_1') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1))$, les différences sont toutes deux des fonctions de l'échantillon de première phase s_1 . Nous savons que X_1 est sans biais pour X_1 et que $B^{(a)}(x^{(a)}; x_1)$ est approximativement sans biais pour $B^{(a)}(x^{(a)}; x_1)$. Sous des conditions assez générales, $N^{-1}(X_1 - X_1') (B^{(a)}(x^{(a)}; x_1) - B^{(a)}(x^{(a)}; x_1)) = O_p(n_1^{-1})$, où n_1 est la taille attendue de s_1 , supposée suffisamment grande. En suivant un raisonnement semblable, $N^{-1}(X - X') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) = O_p(n_1^{-1})$, où n est la taille attendue de s_2 , également supposée suffisamment grande.

Conséquemment, nous pouvons laisser tomber les deux derniers termes de (4.4), parce qu'ils sont d'ordre plus faible que les termes précédents : $N^{-1} \sum^{s_1} a_k (X_i^{(a)} B^{(a)}(x^{(a)}) + e^{(a)}(x^{(a)}; x_1))$ est $O_p(n_1^{-1/2})$ et $N^{-1} \sum^{s_2} a_k e^{(a)}(y_k; x_k)$ est $O_p(n_2^{-1/2})$. Les trois premiers termes définissent la forme linéarisée de $Y_{2p,a}^{2p,w}$.

$$Y_{2p,w}^{2p,w} = \sum^U (X_i^{(a)} B^{(a)}(x^{(a)}) + x_i' B^{(a)}(x^{(a)}; x_1)) + \sum^{s_1} a_{1k}^{(a)} (X_i^{(a)} B^{(a)}(x^{(a)}) + e^{(a)}(x^{(a)}; x_1)) + \sum^S a_k e^{(a)}(y_k; x_k) \quad (4.5)$$

Maintenant, considérons l'expression (4.3) sous le deuxième choix, $a_k^* = w_k a_{2k}$. Cela nous mène à $Y_{2p,w}^{2p,w}$ donné par

$$Y_{2p,w}^{2p,w} = \sum^U (X_i^{(a)} B^{(a)}(x^{(a)}) + x_i' B^{(a)}(x^{(a)}; x_1)) + \sum^{s_1} a_{1k}^{(a)} (X_i^{(a)} B^{(a)}(x^{(a)}) + e^{(a)}(x^{(a)}; x_1)) + \sum^S a_k e^{(a)}(y_k; x_k) + (X_1 - X_1') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) + (X - X') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) \quad (4.6)$$

où $B^{(a)}(x^{(a)}) = (\sum^U w_k a_{2k} z_k x_k^{(a)})^{-1} \sum^U w_k a_{2k} z_k x_k^{(a)}$ et $X = \sum^U w_k a_{2k} x_k^{(a)}$. Les trois premiers termes de $Y_{2p,w}^{2p,w}$ sont les mêmes que ceux figurant dans l'expression (4.4) pour $Y_{2p,a}^{2p,w}$. Les quatrième et cinquième termes diffèrent de leurs homologues dans (4.4). Bien que $B^{(a)}(x^{(a)})$ et X soient des fonctions des poids de calage de première phase w_k , il n'est pas nécessaire que nous les remplacions dans $B^{(a)}(x^{(a)}; x_1)$ et X_1' figurant dans le cinquième terme : cela subdiviserait simplement le terme d'ordre inférieur $(X - X') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1))$ en d'autres termes d'ordre inférieur. Par conséquent, nous pouvons laisser tomber le cinquième terme de (4.6) quand les tailles d'échantillon sont suffisamment grandes. Le quatrième terme peut s'écrire comme il suit.

$$(X_1 - X_1') \left(\sum^{s_1} a_{1k} z_k x_k^{(a)} e^{(a)}(y_k; x_k) \right) = (X_1 - X_1') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) + (X_1 - X_1') (B^{(a)}(x^{(a)}; s_1) - B^{(a)}(x^{(a)}; x_1)) + (X_1 - X_1') \left(\sum^{s_1} a_{1k} z_k x_k^{(a)} \right) - (X_1 - X_1') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) \quad (4.7)$$

Dans cette expression, les quantités sont définies de la façon suivante : $B^{(a)}(x^{(a)}; s_1) = (\sum^{s_1} a_{1k} z_k x_k^{(a)})^{-1} \sum^{s_1} a_{1k} z_k x_k^{(a)}$ et $B^{(a)}(x^{(a)}; s_1)$ ne peut pas être calculée d'après l'échantillon de première phase, parce que les valeurs de y_k ne sont connues que pour $k \in s_2$. Elle est définie implicitement dans le but de déterminer la forme linéarisée. Nous pouvons définir ce genre de construction de la même manière que $B^{(a)}(x^{(a)}; x_1)$ est une fonction de la quantité inconnue $B^{(a)}(x^{(a)}; x_1)$. Or, $B^{(a)}(x^{(a)}; s_1)$ est approximativement sans biais pour sa quantité de population correspondante $B^{(a)}(x^{(a)}; x_1)$. De même, $B^{(a)}(x^{(a)}; s_1)$ est approximativement sans biais pour $\sum^U z_k x_k^{(a)}$. Comme auparavant, nous pouvons soutenir que les trois derniers termes de (4.7) sont d'ordre plus faible que le premier terme $(X_1 - X_1') (B^{(a)}(x^{(a)}; s_1) - B^{(a)}(x^{(a)}; x_1))$, ce qui fournit l'approximation linéaire. L'introduction de ce terme par substitution dans (4.6) mène à la forme linéarisée de $Y_{2p,w}^{2p,w}$.

$$Y_{2p,w}^{2p,w} = \sum^U (X_i^{(a)} B^{(a)}(x^{(a)}) + x_i' B^{(a)}(x^{(a)}; x_1)) + \sum^{s_1} a_{1k}^{(a)} (X_i^{(a)} B^{(a)}(x^{(a)}) + e^{(a)}(x^{(a)}; x_1)) + \sum^S a_k e^{(a)}(y_k; x_k) + (X_1 - X_1') (B^{(a)}(x^{(a)}) - B^{(a)}(x^{(a)}; x_1)) \quad (4.8)$$

[illegible]

4. Comparaison de deux options pour les poids de départ

L'objectif ici est d'analyser comment les poids finaux w_i dans $\hat{Y}^{zp}_i = \sum_j w_j y_j$ dépendent de la spécification des poids de départ a_i dans (3.2). Nous considérons deux cas distincts dépendant du fait que les variables auxiliaires x_i

types d'information auxiliaire dans l'estimation $X = \sum_U y_k$: celui de l'ajustement d'une régression et celui du calage. Sous certaines conditions, ils peuvent aboutir à des estimateurs identiques, mais il n'en est généralement pas ainsi.

L'ajustement d'une régression l'emporte dans Sâmdal et Swensson (1987), Sâmdal, Swensson et Wretman (1992), Stitt (1997), Hidiroglou et Sâmdal (1998), Axelson (1998), ainsi qu'Hidiroglou, Rao et Haziza (2006). L'approche du calage décrite dans Deville et Sâmdal (1992) a été appliquée à l'échantillonnage à deux phases par Dupont (1995). Celle-ci compare les estimateurs par calage résultant à ceux obtenus auxiliaire, les deux approches ne donnent pas nécessairement des estimateurs identiques, mais il est probable qu'en pratique l'écart ait peu de conséquence. Kott et Shukel (1997) ont étudié la méthode du rééchantillonnage pour l'estimation de la variance dans le cas de l'échantillonnage en deux phases. Estévaou et Sâmdal (2002) se concentrent sur l'argument du calage et distinguent dix moyens différents d'utiliser totale-ment ou partiellement l'information disponible aux deux niveaux. Le présent article est également axé sur l'approche par calage. Il étioffe les travaux antérieurs en reconnaissant trois (plutôt que deux) types d'information auxiliaire, possédant chacun des caractéristiques différentes.

Dans l'approche par régression, il est naturel d'ajuster deux régressions par les moindres carrés linéaires. Un ensemble de valeurs de y prédites par régression sont produites pour $k \in s_1$ en utilisant à la fois x_k et x_k^* comme prédicteurs ; un autre est produit pour $k \in s_1$ en utilisant seulement le vecteur x_k comme prédicteur. Les deux ensembles de valeurs prédites de y , ainsi que le total connu $\sum_U x_k$ sont utilisés pour construire l'estimateur de type régression de X , de la façon décrite à la section 9.

L'approche par calage est motivée par deux facteurs : connus ou estimés pour les valeurs auxiliaires et réduite la variance des estimations faites pour la ou les variables étudiées. Nous voulons qu'il y ait une cohérence entre les poids w_k dans $Z_p = \sum_U w_k y_k$ et le total $\sum_U x_k$ connu au niveau de la population et (ou) une estimation (approximativement) sans biais, faite au niveau de l'échantillon de première phase, du total $\sum_U x_k$ inconnu. Puisque y est observé uniquement au dernier niveau (échantillon de deuxième phase), la cohérence « aux niveaux plus élevés » sur les variables auxiliaires importantes réduira souvent de manière significative la variance de $Z_p = \sum_U w_k y_k$. Nous pouvons distinguer deux étapes dans le processus aboutissant aux poids w_k , un calage de première phase et un calage de deuxième phase.

Le plan d'échantillonnage à deux phases est le suivant : de la population finie d'unités $U = \{1, 2, \dots, k, \dots, N\}$, nous tirons un échantillon de première phase s_1 . La

Chaque ensemble peut contenir n'importe quel nombre de variables x . Les trois ensembles sont mutuellement exclusifs. Les propriétés indiquées dans les trois dernières colonnes s'appliquent à chaque variable x dans l'ensemble correspondant. Toutes les variables x utilisées pour le calage appartiennent à l'un de ces trois ensembles.

2. Calage de première phase

Pour le calage de première phase, nous utilisons un vecteur x_k de variables auxiliaires tirées de l'ensemble

Ensemble	Total de la	Valeurs	Valeurs	Ensemble
de variables auxiliaires	variable auxiliaire	variable pour	variable pour	de variables auxiliaires
x^*	sur U	$k \in s_1$	connues	connues
x^*	inconnu	connues	connues	connues
x^*	inconnu	inconnues	inconnues	inconnues

Tableau 1.1. Ensembles de variables auxiliaires pour le calage dans l'échantillonnage à deux phases

concerne l'information.

tableau qui suit donne leurs caractéristiques en ce qui variables x). Ils sont désignés par x^* , x^* et x^* . Le types ou ensembles de variables auxiliaires (appelées tilonnage à deux phases, nous pouvons distinguer trois bilité de l'information auxiliaire. Pour les plans d'échan-estimateurs plus efficaces en tenant compte de la disponi-sans biais pour $X = \sum_U y_k$. Nous pouvons produire des L'estimateur à double facteur d'extension $\sum_U a_k y_k$ est suffisamment grandes.

échantillons de première et de deuxième phases sont l'analyse de nos estimateurs quand les tailles attendues des nous permet d'estimer les termes d'ordre inférieur dans la population et les deux plans d'échantillonnage, ce qui signifié « (approximativement) sans biais par rapport au plan et l'expression » (approximativement) sans biais » présent article, l'analyse des estimateurs est fondée sur le facteur d'extension est $a_k = a_k a_k$ pour $k \in s_1$. Dans le phases s_1 . Le poids de sondage combiné ou à double sont calculés conditionnellement à l'échantillon de première $\pi_{2k|s_1}$ et $a_{2k|s_1}$; il convient de se rappeler que π_{2k} et a_{2k} utilisons π_{2k} et a_{2k} plutôt que les formes plus suggestives $a_{2k} = 1/\pi_{2k}$. (Pour que la notation reste simple, nous poids de sondage conditionnel de deuxième phase est positive de k est $\pi_{2k} = \Pr(k \in s_1)$ pour $k \in s_1$, et le de s_1 . La probabilité d'inclusion conditionnelle connue et sélectionnons un échantillon de deuxième phase s à partir observées pour les unités $k \in s_1$. Alors, sachant s_1 , nous phase est $a_k = 1/\pi_k$. Certaines variables sont peut-être $\pi_k = \Pr(k \in s_1)$ et le poids de sondage de première probabilité d'inclusion positive connue de l'unité k est

Un nouveau visage pour l'échantillonnage à deux phases avec estimateurs par calage

Victor M. Estevao et Carl-Erik Särndal¹

Résumé

Le présent article décrit un cadre pour l'estimation par calage sous les plans d'échantillonnage à deux phases. Les travaux précédents découlaient de la poursuite du développement de logiciels généralisés d'estimation à Statistique Canada. Un objectif important de ce développement est d'offrir une grande gamme d'options en vue d'utiliser efficacement l'information auxiliaire dans différents plans d'échantillonnage. Cet objectif est reflété dans la méthodologie générale pour les plans d'échantillonnage à deux phases exposée dans le présent article.

Nous considérons le plan d'échantillonnage à deux phases classique. Un échantillon de première phase est tiré à partir d'une population finie, puis un échantillon de deuxième phase est tiré en tant que sous-échantillon du premier. La variable étudiée, dont le total de population inconnu doit être estimé, est observée uniquement pour les unités contenues dans l'échantillon de deuxième phase. Des plans d'échantillonnage arbitraires sont permis à chaque phase de l'échantillonnage. Divers types de variables étudiées peuvent être continus ou catégoriques.

L'article apporte une contribution à quatre domaines importants dans le contexte général du calage pour les plans d'échantillonnage à deux phases :

- 1) nous dégageons trois grands types d'information auxiliaire pour les plans à deux phases et les utilisons dans l'estimation. L'information est intégrée dans les poids en deux étapes : un calage de première phase et un calage de deuxième phase. Nous discutons de la composition des vecteurs auxiliaires appropriés pour chaque étape et utilisons une méthode de linéarisation pour arriver aux résidus qui déterminent la variance asymptotique de l'estimateur par calage ;
- 2) nous examinons l'effet de divers choix de poids de départ pour le calage. Les deux choix « naturels » produisent généralement des estimateurs légèrement différents. Cependant, sous certaines conditions, ces deux estimateurs ont la même variance asymptotique ;
- 3) nous réexaminons l'estimation de la variance pour l'estimateur par calage à deux phases. Nous proposons une nouvelle méthode qui peut représenter une amélioration considérable par rapport à la technique habituelle de conditionnement sur l'échantillon de première phase. Une simulation décrite à la section 10 sert à valider les avantages de cette nouvelle méthode ;
- 4) nous comparons l'approche par calage à la méthode de régression assistée par modèle classique qui comporte l'ajustement d'un modèle de régression linéaire à deux niveaux. Nous montrons que l'estimateur assisté par modèle a des propriétés semblables à celles d'un estimateur par calage à deux phases.

Mots clés : Information auxiliaire ; estimateur par la régression à deux phases ; poids de départ ; estimateur de variance à résidus distincts ; estimateur de la variance à résidu combiné.

1. Introduction

L'expression *échantillonnage double* fait référence aux plans d'échantillonnage dont la caractéristique commune est le tirage de deux échantillons probabilistes, désignés par s_1 et s_2 , qui sont tous deux des sous-ensembles de la population finie d'intérêt donnée par $U = \{1, \dots, k, \dots, N\}$. L'échantillon s_1 est réalisé et observé avant l'échantillon s_2 . Une variable étudiée type est désignée par y_k ; sa valeur est obtenue uniquement pour les unités $k \in s_1$. L'objectif est d'estimer le total Y de population $Y = \sum_{k \in U} y_k$ (si A est un ensemble d'unités, $A \subseteq U$, nous utilisons la notation $\sum_{k \in A}$ comme forme abrégée de $\sum_{k \in U}$ s'il n'y a aucune ambiguïté).

Hidiroglou (2001) discute de deux types d'échantillonnage double, *emboîté* et *non emboîté*. Le présent article

porte sur le type *emboîté*, habituellement appelé échantillonnage à deux phases : l'échantillon de deuxième phase s_2 est un sous-échantillon de l'échantillon de première phase s_1 tiré à partir de U , de sorte que $s \subseteq s_1 \subseteq U$. L'estimation sous échantillonnage à deux phases a été examinée par plusieurs auteurs dans le contexte où deux sources d'information auxiliaire sont reconnues et traitées en fonction de leur niveau. Au niveau de la population, le total $\sum_{k \in U} x_k$ est connu, où x_k est un vecteur connu pour chaque $k \in s_1$; par conséquent, il est également connu pour chaque $k \in s_2$. Au niveau du premier échantillon, la valeur du vecteur x_k est observée pour chaque $k \in s_1$ et est par conséquent connue pour chaque $k \in s_2$; le total $\sum_{k \in s_2} x_k$ est inconnu, mais peut être estimé sans biais par rapport au plan au niveau de s_1 . On trouve dans la littérature deux arguments en faveur de l'intégration de ces deux

1. Victor M. Estevao, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6. Courriel : victor.estevao@statcan.gc.ca ; Carl-Erik Särndal, professeur, Courriel : carl.sarndal@rogers.com.

Dans leur article, Schouten, Cobben et Bethlehem se penchent sur l'évaluation de la similité entre la réponse à une enquête et l'échantillon ou la population à l'étude. Ils proposent un indicateur de représentativité pour remplacer les taux de réponse comme indicateur de la qualité lorsque l'on évalue l'effet du biais de non-réponse. Ils montrent que cet indicateur, appelé indicateur R, est apparenté jusqu'à un certain point au V de Cramer qui mesure l'association entre la variable de réponse et les variables auxiliaires. En fait, il est préférable de considérer l'indicateur R comme une mesure de l'absence d'association, puisqu'une association faible implique qu'il n'existe aucune preuve que la non-réponse a eu une incidence sur la composition des données observées. Les auteurs établissent les propriétés théoriques de l'indicateur proposé et illustrent ce dernier au moyen d'études empiriques.

Finalement, dans son article, Chauvet s'attaque au problème de l'échantillonnage équilibré lorsque les tailles dans chaque strate sont trop petites pour permettre un équilibrage exact. L'auteur propose un algorithme adapté de la méthode du Cube garantissant l'équilibrage au niveau de la population. Une étude par simulation confirme la bonne performance de la méthode proposée.

Harold Mantel, Rédacteur en chef délégué

Dans ce numéro

Dans le premier article de ce numéro de *Techniques d'enquête*, Estévaou et Sâmdal examinent le problème de l'estimation par calage dans le contexte de l'échantillonnage à deux phases. Les apports de l'article comprennent le choix des poids initiaux dans la procédure de calage, ainsi que l'important problème de l'estimation de la variance. Les auteurs proposent de nouveaux estimateurs de la variance et procèdent à une étude par simulation, dont les résultats montrent que ces nouveaux estimateurs s'avèrent plus efficaces que les estimateurs classiques.

Ensuite, Li et Valliant étudient le problème de la détection des unités influentes dans l'analyse des données d'enquête par régression linéaire. Ils commencent par donner une expression de la matrice chapeau et des effets de levier comme des éléments diagonaux de la matrice chapeau quand les paramètres du modèle sont estimés par les moindres carrés pondérés. Puis, ils proposent une décomposition des effets de levier et soulignent que l'effet de levier d'une unité donnée peut être important si le poids de sondage de cette dernière est grand ou que son vecteur de variables explicatives est éloigné du centre. Ils illustrent l'effet des unités influentes sur les moindres carrés ordinaires ainsi que pondérés au moyen d'un exemple numérique.

Beaumont et Bocci proposent une méthode bootstrap pour vérifier les hypothèses au sujet d'un vecteur de paramètres inconnus du modèle quand l'échantillon a été tiré d'une population finie. La méthode s'appuie sur des statistiques de test fondées sur un modèle dans lesquelles sont intégrés les poids de sondage et qui peuvent habituellement être obtenues facilement en se servant de logiciels standard. Au moyen d'une étude par simulation, les auteurs montrent que la méthode proposée donne des résultats comparables à ceux de la procédure de Rao-Scott, et de meilleurs résultats que celles de Wald et de Bonferroni lorsque l'on vérifie les hypothèses au sujet d'un vecteur de paramètres d'un modèle de régression linéaire.

Dans leur article, Park, Choi et Choi présentent une approche intéressante de traitement de la non-réponse. Des études ont montré que le comportement de vote des électeurs indécis peut avoir une incidence importante sur le résultat final d'une élection et que l'on peut accroître l'exactitude de la prédiction des résultats en tenant compte de ces électeurs. Les auteurs présentent deux modèles bayésiens dont les distributions de probabilité a priori dépendent de l'information donnée par les répondants et les indécis. Ils analysent un tableau de contingence à double entrée incomplet en se servant de quatre jeux de données provenant des sondages électoraux réalisés en 1998 dans l'État de l'Ohio pour illustrer comment il convient d'utiliser et d'interpréter les résultats des estimations pour les élections.

Ghosh, Kim, Maiti, Katzoff et Parsons élaborent des méthodes bayésiennes hiérarchiques et empiriques pour l'estimation de proportions dans de petits domaines en utilisant des modèles au niveau de l'unité. Ils proposent un modèle hiérarchique bayésien analogue au modèle linéaire mixte généralisé pour obtenir les moyennes et les erreurs-types a posteriori des proportions pour des petits domaines de population. Adoptant une approche fondée sur la théorie des fractions d'estimation optimales, ils obtiennent aussi les estimateurs bayésiens empiriques et les estimateurs asymptotiques de l'erreur quadratique moyenne correspondants. Ils illustrent les méthodes en se servant de données provenant de la National Health Interview Survey (NHIS) pour obtenir des estimations sur petits domaines des proportions d'Asiatiques ne possédant pas d'assurance-maladie.

L'article de McElroy et Holan porte sur les tests de détection de la saisonnalité résiduelle dans les données désaisonnalisées. Les auteurs proposent un test de signification statistique permettant de détecter dans la densité spectrale de la série chronologique étudiée les pics indicateurs de saisonnalité. Ils élaborent la théorie qui sous-tend la méthode proposée et illustrent et comparent cette dernière aux méthodes existantes au moyen d'études par simulation et d'études empiriques.

Gabler et Lahiri fournissent une justification assistée par modèle de la formule classique de la variance due à l'intervieweur pour l'échantillonnage avec probabilités égales sans grappes spatiales. Puis, ils obtiennent, dans le contexte d'un plan d'échantillonnage complexe, une définition de la variabilité due à l'intervieweur qui tient compte comme il convient des probabilités inégales de sélection et des grappes spatiales. Ils proposent aussi une décomposition des effets totaux en effets dus à la pondération, aux grappes spatiales et aux intervieweurs. Leurs résultats peuvent aider à mieux comprendre et contrôler les sources de variabilité.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» — «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 35, numéro 1, juin 2009

Table des matières

Dans ce numéro.....	1
Articles Réguliers	
Victor M. Estevao et Carl-Erik Sämndal Un nouveau visage pour l'échantillonnage à deux phases avec estimateurs par calage.....	3
Jianzhu Li et Richard Valliant Matrice chapeau et effets de levier pondérés par les poids de sondage.....	17
Jean-François Beaumont et Cynthia Bocci Une méthode bootstrap pratique pour les tests d'hypothèses à partir des données d'enquête.....	29
Bo-Seung Choi, Jai Won Choi et Yousung Park Méthodes bayésiennes pour un tableau de contingence à double entrée incomplet avec application aux sondages électoraux de l'Etat de l'Ohio.....	41
Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maji, Myron Katzoff et Van L. Parsons Estimations pour petits domaines bayésiennes hiérarchiques et empiriques de la proportion de personnes sans assurance-maladie dans les groupes de population minoritaires.....	57
Tucker McElroy et Scott Holan Un test non paramétrique pour la saisonnalité résiduelle.....	71
Siegfried Gabler et Partha Lahiri De la définition et de l'interprétation de la variabilité d'intervieweur pour un plan d'échantillonnage complexe.....	91
Barry Schouten, Fannie Cobben et Jelle Bethlehem Indicateurs de la représentativité de la réponse aux enquêtes.....	107
Guillaume Chauvet Échantillonnage équilibré stratifié.....	123

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

D. Royce

Anciens présidents G. J. Brackstone (1986-2005)

R. Plarek (1975-1986)

Membres J. Gambino

R. Jones

J. Kovar

H. Mantel

E. Rancourt

COMITÉ DE RÉDACTION

Rédacteur en chef J. Kovar, *Statistique Canada*

Rédacteur en

H. Mantel, *Statistique Canada*

Rédacteurs associés

J. M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J. L. Elling, *U.S. Bureau of Labor Statistics*

W. A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

M. A. Hidiroglou, *Statistique Canada*

D. Judkins, *Westat Inc.*

D. Kasprzyk, *Mathematical Policy Research*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

S. M. Miller, *Bureau of Labor Statistics*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*

N. G. N. Prasad, *University of Alberta*

A. Zaslavsky, *Harvard University*

C. Wu, *University of Waterloo*

K. M. Wolter, *Iowa State University*

V. J. Verma, *Università degli Studi di Siena*

Y. Tillé, *Université de Neuchâtel*

M. Thompson, *University of Waterloo*

L. Stokes, *Southern Methodist University*

D. Steel, *University of Wollongong*

E. Stasny, *Ohio State University*

P. do N. Silva, *University of Southampton*

F. J. Scheuren, *National Opinion Research Center*

N. Schenker, *National Center for Health Statistics*

L. P. Rivest, *Université Laval*

J. Reiter, *Duke University*

T. J. Rao, *Indian Statistical Institute*

J. N. K. Rao, *Carleton University*

W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou applications à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféralement en Word au rédacteur en chef, re@statcan.gc.ca, Statistique Canada, 150 Promenade du Fré Turney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (6 \$ x 2 exemplaires), autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue
éditée

par Statistique Canada

Juin 2009 • Volume 35 • Numéro 1

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2009

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2009

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa



Statistique
Canada

Statistics
Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements
Service national d'appareils de télécommunications pour les malentendants
Télécopieur
1-800-263-1136
1-800-363-7629
1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements
Télécopieur
1-613-951-8116
1-613-951-0581

Programme des services de dépôt

Service de renseignements
Télécopieur pour le Programme des services de dépôt
1-800-635-7943
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-01-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de choisir la rubrique « Publications ».

Ce produit n° 12-01-X au catalogue est aussi disponible en version imprimée standard au prix de 30 \$CAN. L'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis)
1-800-267-6677
- Télécopieur (Canada et États-Unis)
1-877-287-4369
- Courriel
infostats@statcan.gc.ca
- Poste
Statistique Canada
Finances
Immeuble R.-H.-Coats, 6^e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

Techniques d'enquête

N° 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Juin 2009

•
Volume 35

•
Numéro 1



Statistique
Canada
Statistics
Canada

Canada

12-001

Consultation
Publication

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

December 2009

•

Volume 35

•

Number 2



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and select "Publications."

This product, Catalogue no. 12-001-X, is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail
Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "Providing services to Canadians."

Survey Methodology

A journal
published by
Statistics Canada

December 2009 • Volume 35 • Number 2

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2009

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:
Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 2009

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman J. Kovar

Past Chairmen D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Members J. Gambino
J. Kovar
J. Latimer
H. Mantel
S. Fortier (Production Manager)

EDITORIAL BOARD

Editor J. Kovar, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor M.P. Singh (1975-2005)

Associate Editors

J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
D. Judkins, *Westat Inc*
D. Kasprzyk, *Mathematica Policy Research*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 35, Number 2, December 2009

Contents

In this issue	121
 Waksberg Invited Paper Series	
Graham Kalton Methods for oversampling rare subpopulations in social surveys	125
 Regular Papers	
Andreas Quatember A standardization of randomized response strategies	143
Xiaojian Xu and Pierre Lavallée Treatments for link nonresponse in indirect sampling.....	153
Damião N. da Silva and Jean D. Opsomer Nonparametric propensity weighting for survey nonresponse through local polynomial regression	165
Jan van den Brakel and Sabine Krieg Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design	177
Li-Chun Zhang Estimates for small area compositions subjected to informative missing data	191
Debora F. Souza, Fernando A.S. Moura and Helio S. Migon Small area population prediction via hierarchical models	203
Jun Shao and Katherine J. Thompson Variance estimation in the presence of nonrespondents and certainty strata	215
John Preston Rescaled bootstrap for stratified multistage sampling	227
Donsig Jang and John L. Eltinge Use of within-primary-sample-unit variances to assess the stability of a standard design-based variance estimator.....	235
Zilin Wang and David R. Bellhouse Semiparametric regression model for complex survey data.....	247
Acknowledgements.....	261

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.



In this issue

This issue of *Survey Methodology* opens with the ninth paper in the annual Waksberg Award invited paper series in honour of Joseph Waksberg's contributions to the theory and practice of survey methodology. The editorial board would like to thank the members of the selection committee – Bob Groves, chair, Leyla Mohadjer, Daniel Kasprzyk and Wayne Fuller – for having selected Graham Kalton as the author of this year's Waksberg Award paper.

In his paper entitled "Methods for oversampling rare subpopulations in social surveys" Kalton gives an overview of methods for sampling rare populations, what Kish called minor domains. After discussing general issues he describes several different methods including screening, stratification, two-phase sampling, multiple frames, multiplicity sampling, location sampling, and accumulating samples over time. He discusses the advantages and disadvantages of each method, and gives many examples of their use in surveys. In practice a combination of approaches is often used.

Randomized response strategies are often used in order to reduce nonsampling errors such as nonresponse and measurement errors. They can also be used in the context of statistical disclosure control for public use microdata files. In his paper, Quatember proposes a standardization of randomized response techniques. The statistical properties of the standardized estimator are derived. He applies the proposed method to a survey on academic cheating behaviour.

Xu and Lavallée consider the problem caused by link nonresponse when using the generalized weight share method in indirect sampling. Indirect sampling is used when selecting samples from a population that is not the target population of interest but is related to it. Biased estimates may occur when it is not known that a unit in the sampling population is related to a unit in the target population. The authors propose several weight adjustments to overcome the issue of link nonresponse.

In the context of unit nonresponse, the weights of the respondents are often adjusted by the inverse of the estimated response probability. Da Silva and Opsomer propose to estimate the response probabilities using local polynomial regression. Results of a simulation study are presented confirming the good performance of the proposed method.

In their paper, Van den Brakel and Krieg consider a multivariate structural time series model that accounts for the design of the Dutch Labour Force Survey. The model is used to estimate the unemployment rates. An empirical investigation demonstrates that the proposed model results in a significant increase in accuracy.

Zhang considers estimation of cross-classifications where one margin of the cross-classification corresponds to small areas and where non-response varies from area to area. He develops a double mixed model approach that combines the fixed effects and random area effects of the small area model with the random effects from the missing data mechanism. The associated conditional mean squared error of prediction is approximated in terms of a three-part decomposition, corresponding to a naive prediction variance, a positive correction that accounts for the hypothetical parameter estimation uncertainty based on the latent complete data, and another positive correction for the extra variation due to the missing data.

Souza, Moura and Migon propose a Bayesian small area estimation application using growth models that account for hierarchical and spatial relationships. They use this approach to obtain population predictions for the municipalities not sampled in the Brazilian Annual Household Survey and to increase the precision of the design-based estimates obtained for the sampled municipalities.

Shao and Thompson investigate the problem of variance estimation when a weight adjustment is applied to deal with nonresponse in stratified business surveys. They derive two consistent linearization variance estimators under weak assumptions. Naive jackknife variance estimators do not work well unless the sampling fraction is negligible, which is not the case when there are certainty strata. They propose a modified jackknife variance estimator that is consistent even when there are certainty strata but the non-certainty strata must not have a large sampling fraction. They evaluate their variance estimators empirically using real data and a simulation study.

In his paper, Preston investigates the bootstrap variance estimation for multistage designs when units are selected using simple random sampling without replacement at each stage. He proposes an extension to the commonly used rescaled bootstrap estimator that assumes with replacement sampling or negligible sampling fractions at the first stage. The proposed estimator is compared with the rescaled and Bernoulli bootstrap estimators.

Jang and Eltinge address the problem of estimating degrees of freedom values from stratified multistage designs when a small number of primary sampling units (PSUs) are selected per stratum. Due to the small number of PSUs selected, the traditional Satterthwaite-based degrees of freedom can be a severe underestimate. In their paper, they propose an alternative estimator of the degrees of freedom that uses the within PSU variances to provide auxiliary information on the relative magnitudes of the overall stratum-level variances. The proposed method is illustrated using data from the National Health and Nutrition Examination Survey (NHANES).

The article by Wang and Bellhouse explores an application of nonparametric regression techniques to study the relationship between the response variable and covariates, as well as prediction using auxiliary information in the context of complex surveys. The work is an extension of Bellhouse and Stafford (2001) that used a simple nonparametric regression function to the case of several independent variables, including indicator variables that often appear in regression analysis using survey data.

And finally, we are pleased to inform readers and authors that *Survey Methodology* will shortly be covered by SCOPUS in the Elsevier Bibliographic Databases starting with the June 2008 issue.

Harold Mantel, Deputy Editor

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work. The author receives a cash award made possible by a grant from Westat, in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially by the American Statistical Association.

Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Särndal (2007)
Mary Thompson (2008)
Graham Kalton (2009)
Ivan Fellegi (2010)

Nominations:

The author of the 2011 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, Daniel Kasprzyk, by email to DKasprzyk@Mathematica-Mpr.com. Nominations and suggestions for topics must be received by February 28, 2010.

2009 Waksberg Invited Paper

Author: Graham Kalton

Graham Kalton is Chairman of the Board of Directors and a Senior Vice President at Westat. He has a title of Research Professor in the Joint Program in Survey Methodology at the University of Maryland. Dr. Kalton has wide-ranging interests in survey methodology, and has published on several aspects of the subject, including sample design, nonresponse and imputation, panel surveys, question wording, and coding. He is a Fellow of the American Association for the Advancement of Science, a Fellow of the American Statistical Association, a National Associate of the National Academies, and an elected member of the International Statistical Institute. He delivered the annual Morris Hansen lecture in 2000.

Members of the Waskberg Paper Selection Committee (2009-2010)

Daniel Kasprzyk (Chair), *Mathematica Policy Research*

Wayne A. Fuller, *Iowa State University*

Elizabeth A. Martin

Mary Thompson, *University of Waterloo*

Past Chairs:

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Gordon Brackstone (2005 - 2006)

Sharon Lohr (2006 - 2007)

Robert Groves (2007-2008)

Leyla Mojadjer (2008-2009)

Methods for oversampling rare subpopulations in social surveys

Graham Kalton¹

Abstract

Surveys are frequently required to produce estimates for subpopulations, sometimes for a single subpopulation and sometimes for several subpopulations in addition to the total population. When membership of a rare subpopulation (or domain) can be determined from the sampling frame, selecting the required domain sample size is relatively straightforward. In this case the main issue is the extent of oversampling to employ when survey estimates are required for several domains and for the total population. Sampling and oversampling rare domains whose members cannot be identified in advance present a major challenge. A variety of methods has been used in this situation. In addition to large-scale screening, these methods include disproportionate stratified sampling, two-phase sampling, the use of multiple frames, multiplicity sampling, location sampling, panel surveys, and the use of multi-purpose surveys. This paper illustrates the application of these methods in a range of social surveys.

Key Words: Sample allocation; Screening; Disproportionate stratified sampling; Two-phase sampling; Multiple frames; Location sampling; Panel surveys; Multi-purpose surveys.

1. Introduction

I feel very privileged to have been invited to present this year's paper in the Waksberg Invited Paper Series, a series that honors Joe Waksberg for his numerous contributions to survey methodology. I was extremely fortunate to have had the opportunity to work with Joe at Westat for many years and, as did many others, I benefited greatly from that experience. When faced with an intractable sampling problem, Joe had a flair for turning the problem on its end and producing a workable solution. Since the problem often concerned the sampling of rare populations, I have chosen to review methods for sampling rare populations for this paper.

One of the major developments in survey research over the past several decades has been the continuously escalating demand for estimates for smaller and smaller subclasses (subpopulations) of the general population. This paper focuses on those subclasses – termed *domains* – that are planned for separate analysis at the sample design stage. Some examples of domains that have been taken into account in the sample designs of various surveys include a country's states or provinces, counties or districts; racial/ethnic minorities; households living in poverty; recent births; persons over 80 years of age; recent immigrants; gay men; drug users; and disabled persons. When the domains are small (also known as *rare populations*), the need to provide adequate sample sizes for domain analysis can create major challenges in sample design. This paper reviews the different probability sampling methods that are used to generate samples for estimating the characteristics of rare populations with required levels of precision. Sampling methods for estimating the size of a rare

population are not explicitly addressed, although similar methods are often applicable. However, capture-recapture and related methods are not addressed in this paper.

An important issue for sample design is whether the aim of a survey is to produce estimates for a single domain or many domains. Although much of the literature on the sampling of rare populations discusses sample designs for a single rare domain (e.g., drug users), in practice surveys are often designed to produce estimates for many domains (e.g., each of the provinces in a country or several racial/ethnic groups). The U.S. National Health and Nutrition Examination Survey (NHANES) is an example of a survey designed to produce estimates for many domains, in this case defined by age, sex, race/ethnicity and low-income status (Mohadjer and Curtin 2008). In sample designs that include many domains, the domains may be mutually exclusive (e.g., provinces or the cells of the cross-classification of age group and race/ethnicity) or they may be intersecting (e.g., domains defined separately by age group and by race/ethnicity).

The size of a domain is a key consideration. Kish (1987) proposed a classification of *major domains* of perhaps 10 percent or more of the total population, for which a general sample will usually produce reliable estimates; *minor domains* of 1 to 10 percent, for which the sampling methods in this paper are needed; *mini-domains* of 0.1 to 1 percent, estimates for which mostly require the use of statistical models; and *rare types* comprising less than 0.01 percent of the population, which generally cannot be handled by survey sampling methods. Many surveys aim to produce estimates for some major domains, some minor domains and occasionally even some mini-domains.

1. Graham Kalton, Westat, 1600 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: grahamkalton@westat.com.

Since the sample sizes for most surveys are sufficient to produce estimates of reasonable precision for major domains, there is generally no need to adopt the kinds of oversampling procedures reviewed in this paper. However, there are some important design features that should be considered. It is, for example, valuable to take major domains into account in creating the strata for the survey. This consideration is of particular importance with geographically defined domains and multistage sampling. If a geographic domain is not made into a design stratum, the number of primary sampling units (PSUs) selected in that domain is a random variable; the sampled PSUs in strata that cut across the domain boundaries may or may not be in the domain, creating problems for domain estimation. It is also valuable to have a sizable number of sampled PSUs in each geographical domain in order to be able to compute direct variance estimates of reasonable precision, implying the need to spread the sample across a large number of PSUs. At the estimation stage, it is preferable, where possible, to apply nonresponse and noncoverage post-stratification-type adjustments at the domain rather than the national level. Singh, Gambino and Mantel (1994) and Marker (2001) discuss design issues and Rao (2003, pages 9-25) discusses estimation issues for major domains. Major domains will receive little attention in this paper.

At the other end of the size continuum, even with the use of special probability sampling methods, the sample sizes possible for most surveys are not large enough to produce standard design-based, or direct, estimates of characteristics for multiple domains when many of the domains are mini-domains or rare types. An obvious exception is a national population census, but censuses too have their limitations. Since they are conducted infrequently (in many countries only once a decade), their estimates are dated – a particular concern for mini-domains, which can experience rapid changes. Also, the content of a census must be severely limited in terms of the range of topics and depth of detail. Very large continuous surveys such as the American Community Survey (U.S. Census Bureau 2009a; Citro and Kalton 2007), the French rolling census (Durr 2005) and the German Microcensus (German Federal Statistical Office 2009) have been developed to address the need for more up-to-date data for small domains, but a restriction on content remains (although the content of the German Microcensus does vary over time). Other exceptions occur at the border between mini-domains and minor domains. For example, since 2007 the Canadian Community Health Survey has provided estimates on the health status of the populations of each of Canada's 121 health regions based on an annual survey of around 65,000 persons aged 12 and over, with the production of annual and biennial data files (Statistics Canada 2008). By combining the samples across multiple

years, researchers are able to produce estimates for rare populations of various types.

In general, however, the maximum sample size possible for a survey on a specific topic is not adequate to yield a large set of mini-domain estimates of acceptable precision. Yet policy makers are making increasing demands for local area data at the mini-domain level. This demand for estimates for mini-domains, mainly domains defined at least in part by geographical administrative units, is being addressed by the use of statistical modeling techniques, leading to model-dependent, indirect, small area estimates. Thus, for example, the U.S. Census Bureau's Small Area Income and Poverty Estimates program produces indirect estimates of income and poverty statistics for 3,141 counties and estimates of poor school-age children for around 15,000 school districts every year, based on data now collected in the American Community Survey and predictor variables obtained from other sources available at the local area level, such as tax data (U.S. Census Bureau 2009b). A comprehensive treatment of indirect estimation using small area estimation techniques, a methodology that falls outside the scope of this paper, can be found in Rao (2003).

Apart from location sampling, discussed in Section 3.6, this paper also does not address the various methods that have been developed for sampling other types of mini-domains of much interest to social researchers and epidemiologists, domains that are often "hidden populations" in that the activities defining them are clandestine, such as intravenous drug use (Watters and Biernacki 1989). A range of methods has been developed under the assumption that the members of the mini-domains know each other. The broad class of such designs is termed link-tracing designs (see the review by Thompson and Frank 2000). They are adaptive designs in that the units are selected sequentially, with those selected at later stages dependent on those selected earlier (Thompson and Seber 1996; Thompson 2002).

Snowball sampling was one of the early methods of an adaptive, chain-referral sample design. It starts with some initial sample of rare domain members (the seeds), and they in turn identify other members of the domain. While it bears a resemblance to network (multiplicity) sampling (described in Section 3.5), snowball sampling lacks the probability basis of the latter technique, *i.e.*, known, non-zero, selection probabilities for all members of the domain. A version of snowball sampling has been termed respondent-driven sampling (RDS) (Heckathorn 1997, 2007). Volz and Heckathorn (2008) develop a theory for RDS that is based on four assumptions: (1) that respondents know how many members of the network are linked to them (the degree); (2) that respondents recruit others from their personal network at random; (3) that network connections are reciprocal; and

(4) that recruitment follows a Markov process. The need for these modeling assumptions for statistical inference is the difference between chain-referral sample designs and the conventional probability sample designs used in surveys which do not need to invoke such assumptions. It is apparent that RDS is appropriate only for mini-domains for which clear networks exist. The method is used mainly in local area settings, but Katzoff, Sirken and Thompson (2002) and Katzoff (2004) have suggested that the seeds could come from a large-scale survey, such as the U.S. National Health Interview Survey.

This paper focuses on the use of probability sampling methods to produce standard design-based, or direct, estimates for characteristics of rare populations, building on previous reviews (e.g., Kish 1965a; Kalton and Anderson 1986; Kalton 1993a, 2003; Sudman and Kalton 1986; Sudman, Sirken and Cowan 1988; and Flores Cervantes and Kalton 2008). Much of the literature deals with the sampling issues that arise when the rare population is the sole subject of study. However, as noted above, surveys are often required to produce estimates for many different domains as well as for the total population. Section 2 reviews the design issues involved when the survey has design objectives for multiple domains whose members can be identified from the sampling frame. The main part of the paper, Section 3, provides a review of a range of methods that have been used to sample rare populations whose members cannot be identified in advance. The paper ends with some concluding remarks in Section 4.

2. Multi-domain allocations

The issue of sample allocation arises when a survey is being designed to produce estimates for a number of different domains, for subclasses that cut across the domains, as well as for the total population. In most applications, domains vary considerably in size with at least some of them being rare domains.

Assume that there are H mutually exclusive and exhaustive domains that are identified on the sampling frame. Under the commonly made assumptions that the variance of an estimate for domain h can be expressed as V/n_h and that survey costs are the same across domains, the optimum allocation for estimating the overall population mean is $n_h \propto W_h$, where W_h is the proportion of the population in domain h . Assuming that the domain estimates are all to have the same precision, the optimum allocation is $n_h = n/H$ for all domains. These two allocations are in conflict when the W_h vary greatly, as often occurs when the domains are administrative areas of the country, such as states, provinces, counties or districts. In such cases, adopting the optimum allocation for one

objective leads to a serious loss of precision for the other. However, a compromise allocation that falls between the two optimum allocations often works well for both objectives.

Several compromise solutions exist. One, proposed by Kish (1976, 1988), is to determine the domain sample sizes by the following formula:

$$n_h \propto \sqrt{IW_h^2 + (1 - I)H^{-2}},$$

where I and $(1 - I)$ represent the relative importance of the national estimate and the domain (e.g., administrative district) estimates, respectively. If $I = 1$, the allocation is a proportionate allocation, as optimum for the national estimate, whereas if $I = 0$, the allocation is an equal allocation, as optimum for the domain estimates. The choice of I is highly subjective, but I have found that $I = 0.5$ is often a good starting point, after which a careful review of the allocation can lead to modifications. Bankier (1988) has proposed a similar compromise solution, termed a power allocation. Applied to the current example, the domain sample sizes are determined from $n_h \propto W_h^q$, where q is a power between 0 (equal allocation) and 1 (proportionate allocation). As an example, the 2007 Canadian Community Health Survey was designed to attach about equal importance to the estimates for provinces and health regions. The sample allocation to a province was based on its population size and its number of health regions. Within a province, the sample was allocated between health regions using the Bankier allocation with $q = 0.5$ (Statistics Canada 2008).

A limitation to the Kish and Bankier procedures is that they may not allocate sufficient sample to small domains to produce estimates at the required level of precision. This limitation can be addressed by revising the initial allocations to satisfy precision requirements. An alternative approach addresses this limitation directly: the allocation is determined by fixing a core sample that will satisfy one of the objectives and then supplementing that sample as needed to satisfy the other objective. Singh, Gambino and Mantel (1994) describe such a design for the Canadian Labour Force Survey, with a core sample to provide national and provincial estimates and, where needed, supplemental samples to provide subprovincial estimates of acceptable precision.

The Kish and Bankier schemes assume that the same precision level is required for all small domains. Longford (2006) describes a more general approach in which 'inferential priorities' P_d are assigned to each domain d . As an example, he proposes setting the priorities as $P_d = N_d^a$, where N_d is the population size of domain d and a is a value chosen between 0 and 2. The value $a = 0$ corresponds to the Kish and Bankier equal domain sample size assumption and $a = 2$ corresponds to an overall proportionate

allocation. An intermediate value of a attaches greater priority to larger domains. Longford also extends the approach to incorporate an inferential priority for the overall estimate.

A more general approach to sample allocation is via mathematical programming, as has been proposed by a number of researchers (see, for example, Rodríguez Vera 1982). This approach can accommodate unequal variances across domains, intersecting domains, and multiple estimates for each domain. The U.S. Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) provides an example with intersecting domains, with the sample selected from birth certificate records that contained the requisite domain information. There were 10 domains of interest for the ECLS-B: births classified by race (5 domains), birth weight (3 domains) and twins or non-twins (2 domains). The approach adopted first determined a minimum effective sample size (*i.e.*, the actual sample size divided by the design effect) for each domain. With the 30 cells of the cross-classification of birth weight, race/ethnicity and twin/non-twin treated as strata, an allocation of the sample across the strata was then determined to minimize the overall sample size while satisfying the effective sample size requirements for all the domains (Green 2000).

When there are multiple domains of interest and multi-stage sampling is to be used, a variant of the usual measure of size for probability proportional to size (PPS) sampling can be useful for controlling the sample sizes in the sampled clusters (PSUs, second-stage units, *etc.*), provided that reasonable estimates of the domain population sizes are available by cluster. The requirements that all sampled clusters have approximately the same overall subsample size and that sampled units in each domain have equal probabilities of selection can both be met by sampling the clusters with standard PPS methods, but with a composite measure of size that takes account of the differing sampling rates for different domains (Folsom, Potter and Williams 1987). As an example, in a survey of men in English prisons, the desired sampling fractions were 1 in 2 for civil prisoners (C), 1 in 21 for “star” prisoners who are normally serving their first term of imprisonment (S) and 1 in 45 for recidivists (R). Prisons were selected at the first stage of sampling, with prison i being selected with probability proportional to its composite measure of size $R_i + 2.2S_i + 20.3C_i$, where the multipliers are the sampling rates relative to the rate for recidivists (Morris 1965, pages 303-306).

3. Methods for oversampling rare domains

The main focus of this paper is on the use of probability sampling methods to produce standard design-based, or

direct, estimates for characteristics of rare populations, often minor domains in Kish’s terminology. As preparation for the subsequent discussion, it will be useful to note some features of different types of rare populations that, together with the survey’s mode of data collection, are influential in the choice of sampling methods that can be applied to generate required sample sizes for all domains. Some important features for consideration are summarized below:

- Is a separate frame(s) available for sampling a rare population? Can those sampled be located for data collection? How up-to-date and complete is the frame? If an existing up-to-date frame contains only the rare population (with possibly a few other listings) and provides almost complete coverage, then sampling can follow standard methods. If no single frame gives adequate coverage but there are multiple frames that between them give good coverage, issues of multiple routes of selection arise (Section 3.4).
- Is the rare population concentrated in certain, identifiable parts of the sampling frame, or is it fairly evenly spread throughout the frame? If it is concentrated, disproportionate stratification can be effective (Section 3.2).
- If a sample is selected from a more general population, can a sampled person’s membership in the rare population be determined inexpensively, such as from responses to a few simple questions? If so, standard screening methods may be used (Section 3.1). If accurate determination requires expensive procedures, such as medical examinations, a two-phase design may be useful (Section 3.3). A related issue is whether some members of a rare population consider their membership to be sensitive; the likelihood that members may be tempted to deny their membership may influence the choice of survey administration mode and other aspects of screening.
- Are members of the rare population readily identified by others? If so, some form of network, or multiplicity, sampling may be useful (Section 3.5).
- Are members of the rare population to be found at specific locations or events? If so, location sampling may be useful (Section 3.6).
- Is the rare population defined by a constant characteristic (*e.g.*, race/ethnicity) or by a recent event (*e.g.*, a hospital stay)? The distinction between these two types of characteristics is important in considering the utility of panel surveys for sampling rare populations (Section 3.7).

The following sections review a range of methods for sampling rare populations. Although the methods are discussed individually, some are interrelated and, in practice, a combination of methods is often used.

3.1 Screening

Some form of screening is generally needed when the sampling frame does not contain domain identifiers. This section considers a straightforward application of a screening design in which a large first-phase sample is selected to identify samples of the members of the domains of interest, without recourse to the techniques described in later sections. The first-phase sample size is the minimum sample size that will produce the required (or larger) sample sizes for all of the domains. The minimum first-phase sample size is determined by identifying the required sample size for one of the domains, with all of the sample members of that domain then being included in the second-phase sample. Subsamples of other domains are selected for the second-phase sample at rates that generate the required domain sample sizes. If the survey is designed to collect data for only a subset of the domains (often only one domain), then none of the members of the other domains is selected for the second-phase sample.

Since a very large screening sample size is needed to generate an adequate domain sample size when one (or more) of the domains of interest is a rare population, the cost of screening becomes a major concern. In addition to the sampling methods discussed in later sections, there are several strategies that can be employed to keep costs low:

- Use an inexpensive mode of data collection, such as telephone interviewing or a mail questionnaire, for the screening. The second-phase data collection may be by the same mode or a different mode.
- When possible and useful, permit the collection of screening data from persons other than those sampled. For example, other household members may be able to accurately report the rare population status of the sampled member. See the discussion below and also Section 3.5 on multiplicity sampling.
- When screening is carried out by face-to-face interviewing in a multistage design, it is efficient to select a large sample size in each cluster. Compact clusters can also be used. Costs are reduced, and the precision of domain estimates is not seriously harmed because the average domain sample sizes in the clusters will be relatively small.

One possible means of reducing screening costs is to share the costs across more than one survey. For instance, the child component of the ongoing U.S. National

Immunization Survey (NIS) is a quarterly telephone survey that screens households with landline telephone numbers to locate children aged 19 to 35 months, in order to ascertain vaccination coverage levels (Smith, Battaglia, Huggins, Hoaglin, Roden, Khare, Ezzati-Rice and Wright 2001; U.S. National Center for Health Statistics 2009b). The NIS large-scale screening is also used to identify members of domains of interest for the State and Local Area Integrated Telephone Survey (SLAITS) program, which addresses a variety of other topics over time (U.S. National Center for Health Statistics 2009c). When sharing screening costs across a number of surveys, it is advantageous if the domains for the surveys are fairly disjoint sets in order to minimize the problems associated with screening some respondents into more than one survey.

When no one is at home to complete a face-to-face screening for a household, it may be possible to obtain information from knowledgeable neighbors as to whether the household contains a member of the rare population (e.g., a child under 3 years of age). This approach (which is used in NHANES) can appreciably reduce data collection costs when a large proportion of the households do not contain members of the rare population. However, there is a danger that the approach may result in undercoverage; some protection is provided by requiring that, if the first neighbor interviewed indicates that the household does not include a member of the rare population(s), the other neighbor is also interviewed. Ethical issues also must be considered, particularly for the identification of rare populations that are sensitive in nature.

An extension of the approach of collecting screening information from neighbors is known as focused enumeration. This technique, which is a form of multiplicity sampling (see Section 3.5), involves asking the respondent at each sampled, or “core”, address about the presence of members of the rare population in the n neighboring addresses on either side. In essence, the sample consists of $2n + 1$ addresses for each core address. If the respondent is unable to provide the screening information for one or more of the linked addresses, then the interviewer must make contact at another address. Focused enumeration has been used with $n = 2$ in the British Crime Survey (Bolling, Grant and Sinclair 2008) and the Health Survey of England (Erens, Prior, Korovessis, Calderwood, Brookes and Primatesta 2001) to oversample ethnic minorities. A limitation of the technique is that it will likely produce some (possibly substantial) undercoverage. Evidence of the extent of undercoverage can be obtained by comparing the prevalence of the rare population in the core sample with that in the linked addresses.

In surveys that sample persons by first sampling households, survey designers often prefer to select one person per

household – perhaps allowing two persons to be sampled in large households – to avoid contamination effects and prevent a within-household clustering homogeneity effect on design effects. This design is not always the best (Clark and Steel 2007), and this particularly applies when rare populations are sampled. When rare population members are concentrated in certain households (*e.g.*, minority populations), the size of the screening sample can be appreciably reduced if more than one person – even all eligible persons – can be taken in some households (see Hedges 1973). Elliott, Finch, Klein, Ma, Do, Beckett, Orr and Lurie (2008) suggest that, for oversampling American Indian/Alaskan Native and Chinese minorities in the United States, taking all eligible persons in a household has potential for U.S. health surveys. The NHANES maximizes the number of sampled persons per household. Since each respondent is remunerated for participation, households with more respondents receive more remuneration, a factor thought to increase response rates (Mohadjer and Curtin 2008). Note that within-household homogeneity will have little effect on design effects when the data are analyzed by subgroup characteristics (*e.g.*, age and sex) that cut across households.

The use of large-scale screening to identify rare populations raises three issues, each of which could lead to a failure to achieve planned sample sizes unless precautions are taken. The first results from the fact that, with screening, the sample size for a rare population is a random variable. As a result, the achieved sample size may be larger or smaller than expected. When a minimum sample size is specified for a rare population, it may be wise to determine the sampling fraction to be used to ensure that there is, say, a 90 percent probability that the achieved sample size will be at least as large as the specified minimum. This procedure was used in determining the sampling fractions for the many age, sex and income subdomains for the Continuing Survey of Food Intakes by Individuals 1994-96 (Goldman, Borrud and Berlin 1997).

The second issue raised by large-scale screening is that the overall nonresponse rate must be considered. A sampled member of a rare population will be a nonrespondent if the screener information is not obtained, or if a member of the rare population is identified (perhaps by a proxy informant) but does not respond to the survey items. The overall nonresponse rate may well be much higher than would occur without the screening component. Furthermore, the survey designers must consider the nature of the rare domain and the ways in which members of that domain will react to the survey content. A survey in which new immigrants are asked about their immigration experiences might have a very different response rate than a survey in which war veterans are asked about the medical and other support services they are receiving.

The third issue is that noncoverage can be a significant problem when large-scale screening is used to identify rare populations. One source of noncoverage relates to the sampling frame used for the screener sample. Even though a frame has good overall coverage, its coverage of a rare domain may be inadequate. For example, the noncoverage of a frame of landline telephone numbers is much higher for households of younger people than for the total population. The designers of landline telephone surveys of such rare domains as young children and college students therefore must carefully consider the potential for noncoverage biases. To address the problem of the substantial noncoverage of poor people in telephone surveys, the National Survey of America's Families, which was designed to track the well-being of children and adults in response to welfare reforms, included an area sample of households without telephones in conjunction with the main random digit dialing (RDD) telephone sample (Waksberg, Brick, Shapiro, Flores Cervantes and Bell 1997).

Another source of noncoverage is a failure to identify some members of the rare population at the screening stage. In particular, when a survey aims to collect data only for members of a rare domain, some screening phase respondents may falsely report, and some interviewers may falsely record, that the sampled persons are not members of that domain. These misclassifications may be inadvertent or they may be deliberately aimed at avoiding the second-phase data collection. Misclassification error can give rise to serious levels of noncoverage, particularly when the rare population classification is based on responses to several questions, misreports to any one of which leads to a misclassification (Sudman 1972, 1976). When the survey oversamples one or more rare domains as part of a survey of the general population, misclassifications are uncovered at the second phase, thus avoiding noncoverage. However, misclassifications still result in a smaller sample sizes for rare domains; in addition, the variation in sampling weights between respondents selected as members of the rare domain and those sampled as members of another domain can lead to a serious loss of precision. Noncoverage is more likely to arise when screener data are collected from proxy informants. It is a particular problem with focused enumeration.

In a number of surveys of rare populations, the proportion of rare population members identified has been much lower than prevalence benchmarks. For example, the 1994 NIS had an appreciable shortfall in the identified proportion of children aged 19 to 35 months (4.1 percent compared to the predicted rate of 5 percent) (Camburn and Wright 1996). In the National Longitudinal Survey of Youth of 1997, only 75 percent of youth aged 12 to 23 years were located (Horrigan, Moore, Pedlow and Wolter 1999). These findings could be the result of higher nonresponse rates for

members of the rare population, frame noncoverage of various types, or misclassifications of domain membership. To produce the required sample size, an allowance for under-representation must be made at the design stage. The noncoverage of an age domain appears to be greatest at the domain boundaries, perhaps because respondents do not know exact ages (with those falsely screened out being lost and those falsely screened in being detected and dropped later) or because of deliberate misreporting to avoid the follow-up interview. To counteract this effect, it can be useful to start with an initial screening for all household members or for a broader age range and then narrow down to the required age range later on.

Weighting adjustments can be used in an attempt to mitigate biases caused by nonresponse and noncoverage, but they are necessarily imperfect. Adjustments for a domain specific level of nonresponse require knowledge of the domain membership of nonrespondents, but that is often not available. Adjustments for noncoverage of a rare domain require accurate external data for the domain, data that are often not available. Indeed, one of the purposes for some rare domain surveys is to estimate the domain size. Noncoverage is a major potential source of error in the estimation of domain size.

3.2 Disproportionate stratification

A natural extension of the screening approach is to try to identify strata where the screening will be more productive. In the ideal circumstance, one or more strata that cover all of the rare population and none from outside that population are identified. That case requires no screening process. Otherwise, it is necessary to select samples from all the strata (apart from those known to contain no rare population members) to have complete coverage of the rare population. The use of disproportionate stratification, with higher sampling fractions in the strata where the prevalence of the rare population is higher, can reduce the amount of screening needed.

3.2.1 Theoretical background

Consider initially a survey designed to provide estimates for a single rare population. Waksberg (1973) carried out an early theoretical assessment of the value of disproportionate stratification for this case. Subsequent papers on this topic include those by Kalton and Anderson (1986) and Kalton (1993a, 2003). The theoretical results show that three main factors must be considered in determining the effectiveness of disproportionate stratification for sampling a single rare population: the prevalence rate in each stratum, the proportion of the rare population in each stratum, and the ratio of the full cost of data collection for members of the rare population to the screening cost involved in identifying

members of that population. If it is assumed that (1) the element variances for the rare population are the same across strata and (2) the costs of data collection for members of the rare and non-rare populations are the same across strata, then, with simple random sampling within strata, the optimum sampling fraction in stratum h for minimizing the variance of an estimated mean for the rare population, subject to a fixed total budget, is given by

$$f_h \propto \sqrt{\frac{P_h}{P_h(c-1)+1}},$$

where P_h is the proportion of the units in stratum h that are members of the rare population and c is the ratio of the data collection cost for a sampled member of the rare population to the cost for a member of the non-rare population (Kalton 1993a). The following formula provides the ratio of the variance of the sample mean with the optimum disproportionate stratified sampling fractions to that with a proportionate stratified sample of the same total cost:

$$R = \frac{[\sum A_h \sqrt{P(c-1) + P/P_h}]^2}{P(c-1) + 1},$$

where A_h is the proportion of the rare population in stratum h and P is the prevalence of the rare population in the full population.

In general, the variability in the optimum sampling fractions across the strata, and the gains in precision for the sample mean, decline as c increases. Thus, if the main survey data collection cost is high – as, for instance, when the survey involves an expensive medical examination – or if the screening cost is very low, then disproportionate stratification may yield only minor gains in precision.

When the main data collection cost adds nothing to the screening cost, the ratio of main data collection cost to screening cost will be $c = 1$. In this limiting situation, the formulas given above simplify to $f_h \propto \sqrt{P_h}$ and $R = (\sum \sqrt{A_h W_h})^2$, where W_h is the proportion of the total population in stratum h . These simple formulas provide a useful indication of the maximum variation in optimum sampling fractions and the maximum gains in precision that can be achieved. The square root function in the optimum sampling fraction formula makes clear that the prevalences in the strata must vary a good deal if the sampling fractions are to differ appreciably from a proportionate allocation. For example, even if the prevalence in stratum A is four times as large as that in stratum B, the optimum sampling fraction in stratum A is only twice as large as that in stratum B. The gains in precision ($1 - R$) are large when A_h is large when W_h is small and vice versa. With only two strata, a stratum with a prevalence five times as large as the overall prevalence (i.e., $P_h/P = 5$) will yield gains in precision of 25 percent or more ($(1 - R) \geq 0.25$) only if that stratum

includes at least 60 percent of the rare population (Kalton 2003, Table 1).

In summary, while generally useful, disproportionate stratification will yield substantial gains in efficiency only if three conditions hold: (1) the rare population must be much more prevalent in the oversampled strata; (2) the oversampled strata must contain a high proportion of the rare population; and (3) the cost of the main data collection per sampled unit must not be high. In many cases, not all of these three conditions can be met, in which case the gains will be modest.

Furthermore, the results presented above are based on the assumption that the true prevalence of the rare population in each stratum is known, whereas in practice it will be out of date (for example, based on the last census) or will perhaps simply have been guesstimated. Errors in the prevalence estimates will reduce the precision gains achieved with disproportionate stratification and could even result in a loss of precision. A major overestimation of the prevalence of the rare population, and hence of the optimum sampling fraction, in the high-density stratum can result in a serious loss of precision for the survey estimates. It is therefore often preferable to adopt a conservative strategy, that is, to adopt a somewhat less disproportionate allocation, one that moves in the direction of a proportionate allocation.

3.2.2 Applications

When area sampling is used, data available from the last census and other sources can be used to allocate the area clusters to strata based on their prevalence estimates for the rare population. See Waksberg, Judkins and Massey (1997) for a detailed investigation of this approach for oversampling various racial/ethnic populations and the low-income population using U.S. census blocks and block groups as clusters. Based on data from the 1990 Census, Waksberg and his colleagues found that the approach generally worked well for Blacks and Hispanics but not for the low-income population. While the low-income population did exhibit high concentrations in some blocks and block groups, those areas did not cover a high proportion of that population.

When the survey designers have access to a list frame with names, the names can be used to construct strata of likely members of some racial/ethnic groups. This situation arises, for instance, with lists of names and telephone numbers and when names are merged onto U.S. Postal Service (USPS) Delivery Sequence File addresses (no name merge is made in some cases). The allocation to strata can be based on surnames only or on a combination of surname and first name (and even other names also). Since women often adopt their husbands' surnames, the allocation is generally more effective for men than women. Names can

be reasonably effective for identifying Hispanics, Filipinos, Vietnamese, Japanese and Chinese, but not Blacks. A number of lists of names associated with different racial/ethnic groups have been compiled, such as the list of Spanish names compiled by the U.S. Census Bureau for the 1990s (Word and Perkins 1996). Several commercial vendors have developed complex algorithms to perform racial/ethnic classifications based on names (see Fiscella and Fremont 2006 for further details). The use of names in identifying race and ethnicity has been of considerable interest to epidemiologists and demographers, who have conducted a number of evaluations of this method (*e.g.*, Lauderdale and Kestenbaum 2000; Elliott, Morrison, Fremont, McCaffrey, Pantoja and Lurie 2009). They often assess the effectiveness of the method in terms of positive predictive value and sensitivity, which are the equivalents of prevalence and the proportion of members of the domain who are identified as such by the instrument used for the classification. In the sampling context, besides limitations in the instrument, researchers also need to take into account that sometimes names are not available and that some available names may be incorrect (for example, with address-based sampling, the names may be out-of-date, because the original family has moved out and a new family has moved into an address). These additional considerations serve to reduce the effectiveness of the name stratification, and depending on the particular circumstances, the reduction in effectiveness may be sizable.

As with stratification in general, the stratification factors used for sampling rare populations do not have to be restricted to objective measures. They can equally be subjective classifications. The only consideration is how well they serve the needs of the stratification (see Kish 1965b, pages 412-415, for an example of the effectiveness of the use of listers' rapid classifications of dwellings into low, medium or high socio-economic status for disproportionate stratification). Elliott, McCaffrey, Perlman, Marshall and Hambarsoomians (2009) describe an effective application of subjective stratification for sampling Cambodian immigrants in Long Beach, California. A local community expert rated all individual residences in sampled blocks as likely or unlikely to contain Cambodian households, based on externally observable cultural characteristics such as footwear outside the door and Buddhist altars. The residences allocated to the "likely" stratum (approximately 20 percent) were then sampled at four times the rate than the rest.

Sometimes, when the survey is concerned with producing estimates only for a very rare population, disproportionate stratification may still require an excessive amount of screening. In that circumstance, it may be necessary to sample from the strata where the prevalence is

highest, dropping the other strata and accepting some degree of noncoverage (or redefining the survey population to comprise only members of the rare population in the strata that were sampled). The Hispanic Health and Nutrition Examination Survey of 1982-84 (HHANES) provides an illustration. For its samples of Mexican Americans in the Southwest and Puerto Ricans in the New York City area, the HHANES sampled only from counties with large numbers and/or percentages of Hispanics, based on 1980 Census counts (Gonzalez, Ezzati, White, Massey, Lago and Waksberg 1985).

As another example of this approach, Hedges (1979) describes a procedure for sampling a minority population that is more concentrated in some geographical districts, such as census enumeration districts. In this procedure, the districts are listed in order of their prevalence of members of the rare population (obtained, say, from the last census), and then the survey designers produce Lorenz curves of the cumulative distribution of rare population prevalence and the cumulative distribution of the proportions of rare population members covered. With the cumulative prevalence declining as the cumulative coverage increases, the survey designers can use these distributions to select the combination of prevalence and proportion covered that best fulfills their requirements. The issue then to be faced is whether to make inferences to the covered population, or whether to make inferences to the full population by applying population weighting adjustments in an attempt to address the noncoverage bias.

When a domain is very rare but a portion of it is heavily concentrated in a stratum, researchers sometimes sample that stratum at a rate much higher than the optimum in order to generate a sizable number of cases. Although this approach may produce a large sample of the rare population, the effective sample size (*i.e.*, the sample size divided by the design effect) will be smaller than if the optimum sampling fractions had been used. Thus, from the perspective of the standard survey design-based mode of inference, this approach is not appropriate. However, the researchers using this approach often argue for a model-based mode of inference in which the sampling weights are ignored. In my view, ignoring the sampling weights is problematic. However, discussion of this issue is outside the scope of this paper.

3.3 Two-phase sampling

The screening approach treated in Sections 3.1 and 3.2 assumes that identification of rare population members is relatively easy. When accurate identification is expensive, a two-phase design can be useful, starting with an imperfect screening classification at the first phase, to be followed up with accurate identification for a disproportionate stratified

subsample at the second phase. Whether the two-phase approach is cost-effective depends in part on the relative costs of the imperfect classification and accurate identification: since the imperfect classifications use up some of the study's resources, they must be much less expensive than the accurate identification. Deming (1977) suggests that the ratio of the per-unit costs of the second- to the first-phase data collections should be at least 6:1. Also, the imperfect classification must be reasonably effective in order to gain major benefits from a second-phase disproportionate stratification.

Two- or even three-phase sampling can often be useful in medical surveys of persons with specific health conditions. The first phase of the survey often consists of a screening questionnaire administered by survey interviewers, and the second phase is generally conducted by clinicians, often in a medical center. As one example, in a survey of epilepsy in Copiah County, Mississippi, Haerer, Anderson and Schoenberg (1986) first had survey interviewers administer to all households in the county a questionnaire that had been pretested to ensure that it had a high level of sensitivity for detecting persons with epilepsy. To avoid false negatives at this first phase, a broad screening net was used in identifying persons who would continue to the second phase. All those so identified were the subjects for the second phase of the survey, which consisted of brief neurological examinations conducted by a team of four senior neurologists in a public health clinic.

A second example illustrates the use of another survey to serve as the first-phase data collection for studying a rare domain. In this case, the Health and Retirement Study (HRS) was used as the first phase for a study of dementia and other cognitive impairment in adults aged 70 or older. The HRS collects a wide range of measures on sample respondents, including a battery of cognitive measures. Using these measures, the HRS respondents were allocated to five cognitive strata, with a disproportionate stratified sample being selected for the second phase. The expensive second-phase data collection consisted of a 3- to 4-hour structured in-home assessment by a nurse and neuropsychology technician. The results of the assessment were then evaluated by a geropsychiatrist, a neurologist and a cognitive neuroscientist to assign a preliminary diagnosis for cognitive status, which was then reassessed in the light of data in the person's medical records (Langa, Plassman, Wallace, *et al.* 2005).

A third example is a three-phase design that was used in a pilot study to identify persons who would qualify for disability benefits from the U.S. Social Security Administration if they were to apply for them (Maffeo, Frey and Kalton 2000). At the first phase, a knowledgeable household respondent was asked to provide information about the

disability beneficiary status and impairment status of all adults aged 18 to 69 years in the household. At the second phase, all those classified into a stratum of severely disabled nonbeneficiaries and samples of the other strata were interviewed in person and were then reclassified as necessary into likely disability strata for the third phase. At the third phase, a disproportionate stratified sample of persons was selected to undergo medical examinations in mobile examination centers.

A fairly common practice with two-phase designs is to take no second- (or third-) phase sample from the stratum of those classified as nonmembers of the rare domain based on their responses at the previous phase. The proportion of the population in that stratum is usually very high, and the prevalence of the rare domain in it is very low (indeed, as in the Haerer, Anderson and Schoenberg (1986) study, the stratum is often conservatively defined with the aim of avoiding the inclusion of those who might possibly be members of the rare domain). As a result, a moderate-sized sample from this stratum will yield almost no members of the rare domain. However, the cut-off strategy of taking no sample from this stratum is risky. If the prevalence of the rare domain in this large stratum is more than minimal, a substantial proportion of the domain may go unrepresented in the final sample.

3.4 Multiple frames

Sometimes sampling frames exist that are more targeted on a rare population than a general frame, but they cover only part of the rare population. In this situation, it can be efficient to select the sample from more than one frame. For example, in the common case of oversampling ethnic minorities, there is sometimes a list frame available. The persons on the list can be classified based on their names as being likely to belong to a given ethnic group (*e.g.*, Chinese, Korean, Pacific Islanders, Vietnamese) to create a second, incomplete sampling frame from which to sample, in addition to a more complete frame that has a lower prevalence of the rare population (see, *e.g.*, Elliott *et al.* 2008; Flores Cervantes and Kalton 2008). As with disproportionate stratification (Section 3.2), major benefits derive from this approach only when the second frame has a high prevalence and covers a sizable fraction of the rare population. See Lohr (2009) for a review of the issues involved in sampling from multiple frames.

With multiple frames, some members of the rare population may be included on several frames, in which case they may have multiple routes of being selected into the sample. There are three broad approaches for addressing these multiplicities (Anderson and Kalton 1990; Kalton and Anderson 1986). When all the frames are list frames, as sometimes occurs in health studies, it may be possible to

combine the frames into a single unduplicated list; however, this can often involve difficult record linkage problems. An alternative approach is to make the frames non-overlapping by using a unique identification rule that associates each member of the rare population with only one of the frames, treating the listings on the other frames as blanks (Kish 1965b, pages 388-390). Samples are selected from each of the frames without regard to the duplication, but only the non-blank sampled listings are accepted for the final sample. This approach works best when searches can be made for each sampled unit on the other frames; if the frames are put in a priority order and the unit is found on a prior frame to the one from which the selection was made, the sampled listing would be treated as a blank. In this case, the frames are strata; the sampled units are treated as subclasses within the strata, allowing for the blank listings (Kish 1965b, pages 132-139), and the analysis follows standard methods.

The use of the unique identification approach can, however, be inefficient when the persons sampled from one frame have to be contacted to establish whether their listings are to be treated as real or blank for that frame. In this case, it is generally more economical to collect the survey data for all sampled persons (*i.e.*, to accept the multiple routes of selection). There are, however, exceptions, as in the case of the National Survey of America's Families. That survey used a combination of an area frame and an RDD telephone frame, with the area frame being used to cover only households without telephones (Waksberg, Brick *et al.* 1997). It proved to be efficient to conduct a quick screening exercise with households on the area frame to eliminate households with telephones, retaining only the non-telephone households for the survey.

There are two general approaches for taking multiple routes of selection into account in computing selection probabilities (Bankier 1986; Kalton and Anderson 1986). One method calculates each sampled unit's overall selection probability across all the frames and uses the inverse of that probability as the base weight for the analysis (leading to the Horvitz-Thompson estimator). For example, the overall selection probability for sampled unit i on two frames is $p_i = (p_{1i} + p_{2i} - p_{1i}p_{2i}) = [1 - (1 - p_{1i})(1 - p_{2i})]$, where p_{fi} is the probability of the unit's selection from frame $f = 1, 2$. A variant is to replace the overall selection probability with the expected number of selections (leading to the Hansen-Hurwitz estimator), which is easier to compute when multiple frames are involved. With only two frames, the expected number of selections is $(p_{1i} + p_{2i})$. When selection probabilities are small, there is little difference between these two estimators.

Adjustments to compensate for nonresponse and to calibrate sample totals to known population totals can either be made to the overall selection probabilities p_i or they can

be made to the p_{fi} individually. A problem that can occur is that the survey designers do not know whether a nonresponding unit sampled from one frame is on another frame since that information is only collected in the interview. In this situation the p_i for nonresponding units cannot be directly computed and must be estimated in some fashion. When adjustments are made to the p_{fi} individually, it is not possible to form nonresponse weighting classes that take membership on other frames into account. Instead, the designers must assume that, within weighting classes, the response rates are the same no matter how many frames a unit is on.

In general, the application of the approach described above requires knowledge of each sampled unit's selection probabilities for all of the frames, information that is not always available. When selection probabilities are not known for frames other than the frame(s) from which the unit is sampled (but presence/absence on the frames is known), an alternative approach, termed a weight share method by Lavallée (1995, 2007), can be used. Unbiased estimates of population totals are obtained if the weight for unit i is given by $w_i = \sum_j \alpha_{ij} w'_{ij}$ where α_{ij} are any set of constants such that $\sum_j \alpha_{ij} = 1$ when summed across the j frames, $w'_{ij} = 1/p_{ij}$ if unit i is selected from frame j with probability p_{ij} and $w'_{ij} = 0$ otherwise (Kalton and Brick 1995; Lavallée 2007). For many applications, it is reasonable to set $\alpha_{ij} = \alpha_j$ and then a good choice of α_j is $\alpha_j = \tilde{n}_j / \sum \tilde{n}_j$, where \tilde{n}_j is the effective sample size based on some average design effect (Chu, Brick and Kalton 1999).

The second general approach for dealing with multiple routes of selection uses the multiple-frame methodology introduced by Hartley (1974), and the subject of much recent research (see, e.g., Lohr and Rao 2000 and 2006 and the references cited in those papers). In the case of two frames (A and B), the population can be divided into three mutually exclusive subsets labeled $a = A \cap \bar{B}$, $b = \bar{A} \cap B$ and $ab = A \cap B$. The sample can be divided into samples from a , b and ab , where the ab sample can be separated into respondents sampled from frame A and those sampled from frame B . The samples in subsets a and b have only one route of selection, and hence are readily handled in estimation. Totals for ab could be estimated from the sample from frame A or the sample from frame B , say, \hat{Y}_{ab}^A or \hat{Y}_{ab}^B . The Hartley methodology takes a weighted average of these two estimators, $\hat{Y}_{ab} = \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B$, where θ is chosen to minimize the variance of \hat{Y}_{ab} , taking into account that sample sizes and design effects differ between the two samples. Note that the dual-frame methodology is estimator specific, with different values of θ for different estimators. Skinner (1991), Skinner and Rao (1996) and Lohr and Rao (2006) have proposed an alternative, pseudo-maximum likelihood estimation approach that has the

attraction of avoiding the problems associated with different values of θ for different variables. Wu and Rao (2009) propose a multiplicity-based pseudo empirical likelihood approach for multiple frame surveys, including what they term a single-frame multiplicity-based approach that incorporates Lavallée's weight share method as described above.

When a dual- or multiple-frame design is used, it is often the case that one frame has complete coverage but a low prevalence of the rare population (e.g., an area frame) and the other frame(s) has a much higher prevalence of the rare population but incomplete coverage. Metcalf and Scott (2009), for example, combined an area sample with an electoral roll sample for the Auckland Diabetes, Heart and Health Survey, in which Pacific Islanders, Maoris and older people were domains of special interest. The electoral roll frame had the advantage of containing information about electors' ages, as well as a special roll on which those who considered themselves to be of Maori descent could enroll. Furthermore, many people of Pacific descent could likely be identified by their names, since Pacific languages use fewer letters than English. A disproportionate stratified sample was selected from the electoral roll frame to oversample the domains of interest, and the sample from the area frame brought in people not on the electoral rolls.

The National Incidence Study of Child Abuse and Neglect provides an example of a more complex situation (Winglee, Park, Rust, Liu and Shapiro 2007). That survey used many frames to increase its overall coverage of abused and neglected children. Child Protective Services (CPS) agencies in the sampled PSUs were the basis of the main sampling frame, while police, hospitals, schools, shelters, daycare centers and other agencies were the sources of other frames. The samples from CPS agencies were selected from list frames, but the samples from other agencies were drawn by sampling agencies, constructing rosters of relevant professional staff, and sampling staff who acted as informants about maltreated children. With these procedures, duplication across agencies cannot be ascertained, except in the case of CPS agencies and any of the other agencies. The design was therefore treated as a dual-frame design, with CPS as one frame and the combination of the other frames as the second frame (i.e., assuming no overlap between the other frames).

3.5 Network sampling

Network (or multiplicity) sampling expands on the standard screening approach by asking sampled persons (or addresses) to also serve as proxy informants to provide the screening information for persons who are linked to them in a clearly specified way (Sudman *et al.* 1988; Sirken 2004, 2005). Relatives such as parents, siblings and children are

often used as the basis of linkages. A key requirement is that every member of the linkage must know and be willing to report the rare population membership statuses of all those linked to them. In a pilot study of male Vietnam veterans, Rothbart, Fine and Sudman (1982) included aunts and uncles as informants as well as parents and siblings, but found that aunts and uncles identified far fewer Vietnam veterans than expected. This apparent failure of aunts and uncles to report some veterans gives rise to a potential sampling bias, thus making their inclusion in the linkage rules problematic.

The multiple routes of selection with network sampling need to be taken into account in determining selection probabilities in a similar manner to that described for multiple frames in the previous section. Conceptually, one can consider each member of the rare population divided into, say, l parts corresponding to the l informants for that member; it is then these parts that are sampled for the survey. See Lavallée (2007) for some theory behind the technique.

When network sampling is used in surveys that collect data on the characteristics of rare population members, direct contact must be made with the members of the rare population identified by the initial informant. In this case, the informant has to be able to provide contact information for the rare population members. The linkage definition may be structured to facilitate the follow-up data collection. For example, with face-to-face interviewing, the linkage may be restricted to relatives living in a defined area close to the informant.

Sudman and Freeman (1988) describe the application of network sampling in a telephone survey about access to health care, in which an oversample of persons with a chronic or serious illness was required. During an initial contact with the head of the household, linkages to the respondent's or spouse's parents, stepparents, siblings, grandparents and grandchildren under age 18 were identified and data were collected on their health status. The use of this network sampling design increased the number of chronically or seriously ill adults identified by about one-third. However, about one in eight of the initial network informants with relatives were unable or unwilling to provide illness information for their network members, and 70 percent did not provide complete location information, including 28 percent who provided neither name nor location information (thus making tracing impossible). The use of network sampling led to some false positives (persons reported as being chronically or seriously ill by the initial respondent but reporting themselves as well). A more serious concern is that the survey was not able to provide information on false negatives (this would have required following up a sample of network members reported to be well by the initial informant).

Some forms of linkage have the added benefit that they can incorporate some rare population members who are not on the original sampling frame and would therefore otherwise be a component of noncoverage. For example, Brick (1990) describes a field test for the telephone-based National Household Education Survey (NHES) that used multiplicity sampling to increase the sample of 14- to 21-year-olds, with a focus on school drop-outs. In a subsample of households, all women aged 28 to 65 were asked to provide information for all their 14- to 21-year-old children currently living elsewhere. Some of these children lived in telephone households and hence had two routes of selection. Others lived in non-telephone households and hence would not have been covered by the survey; their inclusion via the multiplicity design increased the coverage rate in 1989 by about 5 percent. However, the response rate for out-of-household youth was much lower than that for in-household youth because of failure to reach the youth, particularly the youth living in non-telephone households.

Tortora, Groves and Peytcheva (2008) provide another example, in this case using multiplicity sampling in an attempt to cover persons with only mobile telephones via an RDD sample of landline telephone numbers. Respondents to the RDD survey (itself a panel survey) were asked to provide information about parents, siblings and adult children living in mobile-only households. The results demonstrate some of the general issues with multiplicity sampling: knowledge about the mobile-only status of the network members depended on the cohesion of the network; there was widespread unwillingness to provide mobile telephone numbers; and many of those identified as mobile-only households in fact also had a landline telephone.

Network sampling has not been widely used in practice for surveys of rare population members. Some of the limitations of the method are illustrated by the studies described above. There is the risk that the sampled informant may not accurately report the rare population status of other members of the linkage, either deliberately or through lack of knowledge. Nonresponse for the main survey data collection is another concern. In addition, ethical issues can arise when sampled persons are asked about the rare population membership of those in their linkage when that membership is a sensitive matter. The benefits of network sampling are partially offset by the increased sampling errors arising from the variable weights that the method entails, and by the costs of locating the linked rare population members.

3.6 Location sampling

Location sampling is widely used to sample populations that have no fixed abode for both censuses and surveys: nomads may, for example, be sampled at waterpoints when

they take their animals for water, and homeless persons may be sampled at soup kitchens when they go for food (e.g., Kalton 1993a; Ardilly and Le Blanc 2001). A central feature of such uses of location sampling is that there is a time period involved, resulting in issues of multiplicity (Kalsbeek 2003). A serious concern with the use of the technique is that it fails to cover those who do not visit any of the specified locations in the particular time period.

Location sampling is used to sample rare mobile populations such as passengers at airports and visitors to a museum or national park. In such cases, the question arises as to whether the unit of analysis should be the visit or the visitor. When the visit is the appropriate unit, no issues of multiplicity arise (see, for example, the report on the U.S. National Hospital Discharge Survey by DeFrances, Lucas, Buie and Golosinskiy 2008). However, when the visitor is the unit of analysis, the fact that visitors may make multiple visits during the given time period must be taken into account (Kalton 1991; Sudman and Kalton 1986). One approach is to treat visits as eligible only if they are the first visits made during the time period for the survey. Another approach is to make multiplicity adjustments to the weights in the analysis; however, determining the number of visits made is problematic because some visits will occur after the sampled visit.

Location sampling has also been used for sampling a variety of rare – often very rare – populations that tend to congregate in certain places. For example, Kanouse, Berry and Duan (1999) employed the technique to sample street prostitutes in Los Angeles County by sampling locations where street prostitution was known to occur, and by sampling time periods (days and shifts within days). Location (center) sampling has also been used to sample legal and illegal immigrants in Italy (Meccati 2004). For a 2002 survey of the immigrant population of Milan, 13 types of centers were identified, ranging from centers that provide partial lists from administrative sources (e.g., legal and work centers, language courses), centers that have counts of those attending (e.g., welfare service centers, cultural associations), to centers with no frame information (e.g., malls, ethnic shops).

Location sampling has often been used to sample men who have sex with men, with the locations being venues that such men frequent, such as gay bars, bathhouses and bookstores (Kalton 1993b, MacKellar, Valleroy, Karon, Lemp and Janssen 1996). Based on a cross-sectional telephone survey, Xia, Tholandi, Osmond, Pollack, Zhou, Ruiz and Catania (2006) found that men who visited gay venues more frequently had higher rates of high-risk sexual behaviors and also that the rates of high-risk behaviors varied by venue. These findings draw attention to the difficulty of generating a representative sample by location sampling.

McKenzie and Mistiaen (2009) carried out an experiment to compare location (intercept) sampling with both area sampling and snowball techniques, for sampling Brazilians of Japanese descent (Nikkei) in Sao Paulo and Parana. The locations included places where the Nikkei often went (e.g., a sports club, a metro station, grocery stores and a Japanese cultural club) and events (e.g., a Japanese film and a Japanese food festival). Based on this experiment, they conclude that location sampling (and snowball sampling) oversampled persons more closely connected with the Nikkei community and thus did not produce representative samples. This not-unexpected finding highlights the concern about the use of location sampling for sampling rare populations in general, although not for sampling visits to specified sites.

3.7 Accumulating or retaining samples over time

When survey data collection is repeated over time, survey designers can take advantage of that feature in sampling rare populations (Kish 1999). An important distinction to be made is that between repeated and panel surveys. Samples of rare population members can readily be accumulated over time in repeated surveys. For example, the U.S. National Health Interview Survey is conducted on a weekly basis with nationally representative samples; samples of rare populations can be accumulated over one or more years until a sufficient sample size is achieved (U.S. National Center for Health Statistics 2009a). With accumulation over time, the estimates produced are period, rather than point-in-time, estimates that can be difficult to interpret when the characteristics of analytic interest vary markedly over time (Citro and Kalton 2007). For example, how is a 3-year period poverty rate for a rare minority population to be interpreted when the poverty rate has varied a great deal over the period?

In considering the sampling of rare populations in panel surveys, it is important to distinguish between rare populations that are defined by static versus non-static characteristics. No accumulation over time can be achieved in panel surveys for rare populations defined by static characteristics such as race/ethnicity. However, if a sample of a static rare population is taken at one point in time, it can be useful to follow that sample in a panel to study that population's characteristics at later time points, possibly with supplementary samples added to represent those who entered that population after the original sample was selected. Fecso, Baskin, Chu, Gray, Kalton and Phelps (2007) describe how this approach has been applied in sampling U.S. scientists and engineers over a decade. For the decade of the 1990s, the National Survey of College Graduates (NSCG) was conducted in 1993 with a stratified sample of college graduates selected from the 1990 Census of Population long-form sample records. Those found to be

scientists or engineers were then resurveyed in the NSCG in 1995, 1997 and 1999. To represent new entrants to the target population, another survey – the Survey of Recent College Graduates – was conducted in the same years as the NSCG. A subsample of the recent college graduates was added in to the next round of the NSCG panel on each occasion.

Panel surveys can be used to accumulate samples of non-static rare populations, especially persons experiencing an event such as a birth or a divorce. The U.S. National Children's Study, for instance, plans to follow a large sample of eligible women of child-bearing age over a period of about four years, enrolling those who become pregnant in the main study, a longitudinal study that will follow the children through to age 21 (National Children's Study 2007, Michael and O'Muircheartaigh 2008).

Finally, a large sample can be recruited into a panel and provide data that will identify members of a variety of rare populations that may be of future interest. They are then followed in the panel and, based on their rare population memberships, included in the samples for the surveys for which they qualify. Körner and Nimmergut (2004) describe a German "access panel" that could be used in this way, and there are now several probability-based Web panels that can serve this purpose (Callegaro and DiSogra 2008). However, a serious concern with such panels is the low response rates that are generally achieved.

4. Concluding remarks

This paper has presented a brief overview of the range of methods used in sample surveys for sampling and oversampling rare populations, primarily those classified by Kish as minor domains (the references cited provide more details). Although the methods have been discussed separately, in practice they are often combined, particularly when there are several rare domains of interest. As an example, the California Health Interview Survey, conducted by telephone, has used a combination of disproportionate stratification (oversampling telephone exchanges where the prevalence of the Korean and Vietnamese populations of interest is higher) and a dual-frame design (RDD methods supplemented with a frame of likely Korean and Vietnamese names). In many cases, the art of constructing an effective probability sample design for a rare population is to apply some combination of methods in a creative fashion.

As another example, the Pew Research Center telephone survey of Muslim Americans employed three sampling methods to sample this very rare population (Pew Research Center 2007). One component of the design was a geographically stratified RDD sample, with disproportionate stratified

sampling from strata defined in terms of the prevalence of Muslim Americans. The stratum with the lowest prevalence was treated as a cut-off stratum and excluded. The second component was a recontact sample of Muslim Americans drawn from Pew's interview database of recent surveys. The third component was an RDD sample selected from a list of likely Muslim Americans provided by a commercial vendor. To avoid duplicate routes of selection between the geographical strata and the commercial vendor list, telephone numbers selected from the geographical strata were matched against the commercial vendor list and dropped from the geographical strata sample if a match was found.

Not only are the various sampling techniques often used in combination in sample designs for rare populations, but several of the techniques are interrelated. For example, multiple frames can be treated by unique identification (see Section 3.4), which in effect is simply disproportionate stratification. Whereas the whole population is classified into strata for disproportionate stratification, the same approach is adopted with two-phase sampling, but the classification into strata is applied only to members of the first-phase sample. The theory of network sampling is similar to that of multiple-frame sampling, when the latter technique uses inverse overall selection probabilities as weights in the analysis. These interrelationships help to explain the similarities in the theoretical underpinnings of the techniques.

Acknowledgements

I would like to thank Daniel Levine and Leyla Mohadjer for helpful reviews of a draft of this paper, Daifeng Han and Amy Lin for constructive comments on an earlier, shorter version of the paper, and to Mike Brick, Marc Elliott, and Jon Rao for advice on some specific points.

References

- Anderson, D.W., and Kalton, G. (1990). Case-finding strategies for studying rare chronic diseases. *Statistica Applicata*, 2, 309-321.
- Ardilly, P., and Le Blanc, D. (2001). Sampling and weighting a survey of homeless persons: A French example. *Survey Methodology*, 27, 109-118.
- Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *American Statistician*, 42, 174-177.
- Bolling, K., Grant, C. and Sinclair, P. (2008). *2006-07 British Crime Survey (England and Wales)*. Technical Report. Volume I. Available at <http://www.homeoffice.gov.uk/rds/pdfs07/bcs0607tech1.pdf>.

- Brick, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.
- Callegaro, M., and DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 1008-1032.
- Camburn, D.P., and Wright, R.A. (1996). Predicting eligibility rates for rare populations in RDD screening surveys. Available at http://www.cdc.gov/nis/pdfs/sample_design/camburn1996.pdf.
- Chu, A., Brick, J.M. and Kalton, G. (1999). Weights for combining surveys across time or space. *Bulletin of the International Statistical Institute, Contributed Papers*, 2, 103-104.
- Citro, C.F., and Kalton, G. (Eds.) (2007). *Using the American Community Survey: Benefits and Challenges*. Washington, DC: National Academies Press.
- Clark, R.G., and Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society*, 170, Series A, 63-82.
- DeFrances, C.J., Lucas, C.A., Buie, V.C. and Golosinskiy, A. (2008). *2006 National Hospital Discharge Survey*. National Health Statistics Reports Number 5. U.S. National Center for Health Statistics, Hyattsville, MD.
- Deming, W.E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *International Statistical Review*, 45, 29-37.
- Durr, J.-M. (2005). The French new rolling census. *Statistical Journal of the United Nations Economic Commission for Europe*, 22, 3-12.
- Elliott, M.N., Finch, B.K., Klein, D., Ma, S., Do, D.P., Beckett, M.K., Orr, N. and Lurie, N. (2008). Sample designs for measuring the health of small racial/ethnic subgroups. *Statistics in Medicine*, 27, 4016-4029.
- Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantoja, P. and Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services Outcomes Research Methods*, 9, 69-83.
- Elliott, M.N., McCaffrey, D., Perlman, J., Marshall, G.N. and Hambarsoomians, K. (2009). Use of expert ratings as sampling strata for a more cost-effective probability sample of a rare population. *Public Opinion Quarterly*, 73, 56-73.
- Erens, B., Prior, G., Korovessis, C., Calderwood, L., Brookes, M. and Primatesta, P. (2001). Survey methodology and response. In *Health Survey for England – The Health of Minority Ethnic Groups '99. Volume 2: Methodology and Documentation*. (Eds., B. Erens, P. Primatesta and G. Prior). The Stationery Office, London.
- Fecso, R.S., Baskin, R., Chu, A., Gray, C., Kalton, G. and Phelps, R. (2007). *Design Options for SESTAT for the Current Decade*. Working Paper SRS 07-021. Division of Science Resource Statistics, U.S. National Science Foundation.
- Fiscella, K., and Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41, 1482-1500.
- Flores Cervantes, I., and Kalton, G. (2008). Methods for sampling rare populations in telephone surveys. In *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japac, P. Lavrakas, M.W. Link and R.L. Sangster). Hoboken, NJ: Wiley, 113-132.
- Folsom, R.E., Potter, F.J. and Williams, S.K. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 792-796.
- German Federal Statistical Office (2009). *Microcensus*. Available at http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/EN/press/abisz/Mikrocensus__e,templateId=renderPrint.psm.
- Goldman, J.D., Borrud, L.G. and Berlin, M. (1997). An overview of the USDA's 1994-96 Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 796-801.
- Gonzalez, J.F., Ezzati, T.M., White, A.A., Massey, J.T., Lago, J. and Waksberg, J. (1985). Sample design and estimation procedures. In *Plan and Operation of the Hispanic Health and Nutrition Examination Survey, 1982-84*. (Ed., K.R. Maurer). Vital and Health Statistics, Series 1, No. 19. U.S. Government Printing Office, Washington, DC, 23-32.
- Green, J. (2000). Mathematical programming for sample design and allocation problems. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 688-692.
- Haerer, A.F., Anderson, D.W. and Schoenberg, B.S. (1986). Prevalence and clinical features of epilepsy in a biracial United States population. *Epilepsia*, 27, 66-75.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, 36, 99-118.
- Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
- Heckathorn, D.D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37, 151-208.
- Hedges, B.M. (1973). *Sampling Minority Groups*. Thomson Medal Awards. Thomson Organization, London.
- Hedges, B.M., (1979). Sampling minority populations. In *Social and Educational Research in Action* (Ed., M.J. Wilson) London: Longman, 245-261.
- Horrigan, M., Moore, W., Pedlow, S. and Wolter, K. (1999). Undercoverage in a large national screening survey for youths. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 570-575.
- Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
- Kalton, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183-194.
- Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, United Nations, New York.

- Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. In *Methodological Issues in AIDS Behavioral Research*. (Eds., D.G. Ostrow, R.C. Kessler). New York: Plenum Press.
- Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition*, 6, 491-501.
- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, 65-82.
- Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- Kanouse, D.E., Berry, S.H. and Duan, N. (1999). Drawing a probability sample of female street prostitutes in Los Angeles County. *Journal of Sex Research*, 36, 45-51.
- Katzoff, M.J. (2004). Applications of adaptive sampling procedures to problems in public health. In *Proceedings of Statistics Canada Symposium 2004, Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8751-eng.pdf>
- Katzoff, M.J., Sirken, M.G. and Thompson, S.K. (2002). Proposals for adaptive and link-tracing sampling designs in health surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1772-1775.
- Kish, L. (1965a). Selection techniques for rare traits. In *Genetics and the Epidemiology of Chronic Diseases*. Public Health Service Publication No. 1163.
- Kish, L. (1965b). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L. (1976). Optima and proxima in linear sample design. *Journal of the Royal Statistical Society, A*, 139, 80-95.
- Kish, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.
- Kish, L. (1988). Multipurpose sample design. *Survey Methodology*, 14, 19-32.
- Kish, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, 25, 129-138.
- Kömer, T., and Nimmergut, A. (2004). A permanent sample as a sampling frame for difficult-to-reach populations? In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8752-eng.pdf>.
- Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter, G.G., Rodgers, W.L., Steffens, D.C., Weir, D.R. and Willis, R.J. (2005). The Aging, Demographics, and Memory Study: Study design and methods. *Neuroepidemiology*, 25, 181-191.
- Lauderdale, D.S., and Kestenbaum, B. (2000). Asian American ethnic identification by surname. *Population Research and Policy Review*, 19, 283-300.
- Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.
- Lohr, S.L. (2009). Multiple-frame surveys. In *Handbook of Statistics. Volume 29A: Sample Surveys: Design, Methods, and Applications*. (Eds., D. Pfeffermann and C.R. Rao). Burlington, MA: Elsevier B.V.
- Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.L., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.
- MacKellar, D., Valleroy, L., Karon, J., Lemp, G. and Janssen, R. (1996). The Young Men's Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111, Supplement 1, 138-144.
- Maffeo, C., Frey, W. and Kalton, G. (2000). Survey design and data collection in the Disability Evaluation Study. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 79-88.
- Marker, D.A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.
- McKenzie, D.J., and Mistiaen, J. (2009). Surveying migrant households: A comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society*, 172, 339-360.
- Meccati, F. (2004). Center sampling: A strategy for sampling difficult-to-sample populations. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8740-eng.pdf>.
- Metcalf, P., and Scott, A. (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28, 1512-1523.
- Michael, R.T., and O'Muircheartaigh, C.A. (2008). Design priorities and disciplinary perspectives: The case of the US National Children's Study. *Journal of the Royal Statistical Society, A*, 171, 465-480.
- Mohadjer, L., and Curtin, L.R. (2008). Balancing sample design goals for the National Health and Nutrition Examination Survey. *Survey Methodology*, 34, 119-126.
- Morris, P. (1965). *Prisoners and Their Families*. Allen and Unwin, London, 303-306.
- National Children's Study (2007). Study design. In *The National Children's Study Research Plan*. Version 1.3. Available at http://www.nationalchildrensstudy.gov/research/studydesign/researchplan/Pages/Chapter_6_032008.pdf
- Pew Research Center (2007). *Muslim Americans, Middle Class and Mostly Mainstream*. Pew Research Center, Washington, DC.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rodriguez Vera, A. (1982). Multipurpose optimal sample allocation using mathematical programming. Biostatistics Doctoral Dissertation. Ann Arbor: University of Michigan.

- Rothbart, G.S., Fine, M. and Sudman, S. (1982). On finding and interviewing the needles in the haystack: The use of multiplicity sampling. *Public Opinion Quarterly*, 46, 408-421.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.
- Sirken, M.G. (2004). Network sample surveys of rare and elusive populations: A historical review. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8614-eng.pdf>.
- Sirken, M.G. (2005). Network sampling developments in survey research during the past 40+ years. *Survey Research*, 36, 1, 1-5. Available at <http://www.srl.uic.edu/Publist/Newsletter/pastissues.htm>.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Smith, P.J., Battaglia, M.P., Huggins, V.J., Hoaglin, D.C., Roden, A., Khare, M., Ezzati-Rice, T.M. and Wright, R.A. (2001). Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine*, 20(4S), 17-24.
- Statistics Canada (2008). *Canadian Community Health Survey (CCHS)*. Available at <http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&lang=en&db=imdb&adm=8&dis=2#b3>.
- Sudman, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association*, 67, 335-339.
- Sudman, S. (1976). *Applied Sampling*. New York: Academic Press.
- Sudman, S., and Freeman, H.E. (1988). The use of network sampling for locating the seriously ill. *Medical Care*, 26, 992-999.
- Sudman, S., and Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.
- Sudman, S., Sirken, M.G. and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.
- Thompson, S.K. (2002). *Sampling*. 2nd Edition. New York: John Wiley & Sons, Inc.
- Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.
- Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.
- Tortora, R., Groves, R.M. and Peytcheva, E. (2008). Multiplicity-based sampling for the mobile telephone population: Coverage, nonresponse, and measurement issues. In *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link and R.L. Sangster). Hoboken, NJ: Wiley, 133-148.
- U.S. Census Bureau (2009a). *Design and Methodology, American Community Survey*. U.S. Government Printing Office, Washington, DC.
- U.S. Census Bureau (2009b). *Small Area Income and Poverty Estimates*. Available at <http://www.census.gov/did/www/saipe/methods/statecounty/index.html>.
- U.S. National Center for Health Statistics (2009a). *National Health Interview Survey (NHIS)*. Available at <http://www.cdc.gov/nchs/nhis/methods.htm>.
- U.S. National Center for Health Statistics (2009b). *The National Immunization Survey (NIS)*. Available at http://www.cdc.gov/nis/about_eng.htm.
- U.S. National Center for Health Statistics (2009c). *State and Local Area Integrated Telephone Survey (SLAITS)*. Available at <http://www.cdc.gov/nchs/about/major/slaits/nsch.htm>.
- Volz, E., and Heckathorn, D.D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24, 79-97.
- Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section*, American Statistical Association, 429-434.
- Waksberg, J., Brick, J.M., Shapiro, G., Flores Cervantes, I. and Bell, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 713-718.
- Waksberg, J., Judkins, D. and Massey, J.T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61-71.
- Watters, J.K., and Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.
- Winglee, M., Park, I., Rust, K., Liu, B. and Shapiro, G. (2007). A case study in dual-frame estimation methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3195-3202.
- Word, D.L., and Perkins, R.C. (1996). *Building a Spanish Surname List for the 1990's - A New Approach to an Old Problem*. Population Division Technical Working Paper No. 13. U.S. Census Bureau, Washington, DC.
- Wu, C., and Rao, J.N.K. (2009). Empirical likelihood methods for inference from multiple frame surveys. *Proceedings of the International Statistical Institute*, Durban, South Africa.
- Xia, Q., Tholandi, M., Osmond, D.H., Pollack, L.M., Zhou, W., Ruiz, J.D. and Catania, J.A. (2006). The effect of venue sampling on estimates of HIV prevalence and sexual risk behaviors in men who have sex with men. *Sexually Transmitted Diseases*, 33, 545-550.

A standardization of randomized response strategies

Andreas Quatember¹

Abstract

Randomized response strategies, which have originally been developed as statistical methods to reduce nonresponse as well as untruthful answering, can also be applied in the field of statistical disclosure control for public use microdata files. In this paper a standardization of randomized response techniques for the estimation of proportions of identifying or sensitive attributes is presented. The statistical properties of the standardized estimator are derived for general probability sampling. In order to analyse the effect of different choices of the method's implicit "design parameters" on the performance of the estimator we have to include measures of privacy protection in our considerations. These yield variance-optimum design parameters given a certain level of privacy protection. To this end the variables have to be classified into different categories of sensitivity. A real-data example applies the technique in a survey on academic cheating behaviour.

Key Words: Privacy protection; Statistical disclosure control; Nonresponse; Untruthful answering.

1. Introduction

The occurrence of nonresponse and the unwillingness to provide the true answers are natural in survey sampling. They may result in an estimator of population parameters, which has a bias of unknown magnitude and a high variance. A responsible user therefore cannot ignore the presence of nonresponse and untruthful answering.

Let U be the universe of N population units and U_A be a subset of N_A elements, that belong to a class A of a categorical variable under study. Moreover let U_A^c be the group of N_A^c elements, that do not belong to this class ($U = U_A \cup U_A^c$, $U_A \cap U_A^c = \emptyset$, $N = N_A + N_A^c$). Let

$$x_i = \begin{cases} 1 & \text{if unit } i \in U_A, \\ 0 & \text{otherwise} \end{cases}$$

($i = 1, 2, \dots, N$) and the parameter of interest be the relative size π_A of subpopulation U_A :

$$\pi_A = \frac{\sum_U x_i}{N} = \frac{N_A}{N} \quad (1)$$

($\sum_U x_i$ is abbreviated notation for $\sum_{i \in U} x_i$). In a probability sample s (see for instance: Särndal, Swensson and Wretman 1992, page 8f) an estimator of π_A can be calculated from the Horvitz-Thompson estimator of N_A by

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \sum_s \frac{x_i}{\pi_i} \quad (2)$$

($\pi_i > 0$ is the probability that unit i will be included in the sample), if the question "Are you a member of group U_A ?" (or an equivalent question) is asked directly (dir). This estimator is unbiased, if all x_i 's ($i = 1, 2, \dots, n$) are

observed truthfully. In the presence of unit or item nonresponse with respect to a variable under study the sample s is divided into a "response set" $r \subset s$ of size n_r and a "missing set" $m \subset s$ of size n_m ($s = r \cup m$, $r \cap m = \emptyset$, $n = n_r + n_m$). For variables of a highly personal, embarrassing matter (like drug addiction, diseases, sexual behaviour, tax evasion, alcoholism, domestic violence or involvement in crimes) r is furthermore divided into a set t of n_t sample units, who answer truthfully, and a set u of size n_u , who answer untruthfully ($r = t \cup u$, $t \cap u = \emptyset$, $n_r = n_t + n_u$). Estimator (2) must then be rewritten as:

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \left(\sum_t \frac{x_i}{\pi_i} + \sum_u \frac{x_i}{\pi_i} + \sum_m \frac{x_i}{\pi_i} \right). \quad (3)$$

Evidently the elements of set u cannot be identified and the x_i 's of m are not observable. This imposes errors of measurement and nonresponse on the estimation. Therefore everything should be done to keep the untruthful answering rate as well as the nonresponse rate as low as possible.

Survey design features, which clearly affect both the quantity and the quality of the information asked from the respondents (see for instance: Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004, Section 6.7), are strongly related to the sample units' concerns about "data confidentiality" and "perceived protection of privacy". The first term refers to the respondents' desire to keep replies out of hands of uninvolved persons, whereas the second refers to the wish to withhold information from absolutely anybody. Singer, Mathiowetz and Couper (1993) and Singer, van Hoewyk and Neugebauer (2003) report on two successive U.S. population surveys, that the higher these concerns are the lower is the probability of the respondent's participation in the survey (page 470ff and page 375ff).

1. Andreas Quatember is Assistant Professor at the IFAS-Department of Applied Statistics, Johannes Kepler University Linz, Altenberger Str. 69, A-4040 Linz, Austria, Europe. Web address: www.ifas.jku.at. E-mail: andreas.quatember@jku.at.

What can statisticians contribute to this important field of research? For awkward questions the use of *randomized response strategies* at the survey's design stage may reduce the rates of nonresponse and of untruthful answering due to a perceived increase of privacy protection. A common characteristic of these methods is that instead of the direct questioning on the sensitive subject a questioning design is used, which does not enable the data collector to identify the (randomly selected) question on which the respondent has given the answer, although it does still allow to estimate the parameter under study. The idea is to reduce in this way the individuals' fear of an embarrassing "outing" to make sure that the responding person is willing to cooperate. To achieve this goal the respondent clearly has to understand how the questioning design does protect his or her privacy (cf. Landsheer, van der Heijden and van Gils 1999, page 6ff).

Pioneering work in this field was published by Warner (1965). In his questioning design each respondent has to answer randomly either with probability p_1 the question "Are you a member of group U_A ?" or with probability $p_2 = 1 - p_1$ the alternative "Are you a member of group U_A^c ?" ($0 < p_1 < 1$). Since then various randomized response techniques with differing randomization devices have been proposed (for a review see: Chaudhuri and Mukerjee 1987, Nathan 1988 or Tracy and Mangat 1996). All of these strategies make use of randomly selected questions or answers, though some of them use different random devices depending on the respondent's possession or nonpossession of a certain attribute (see for example: Kuk 1990; Mangat 1994; Kim and Warde 2005).

Warner (1971) was the first to note that these techniques are also applicable as methods of masking confidential micro-data sets to allow their release for public use (cf. ibd., page 887). Such microdata sets might contain variables, which allow the direct identification of survey units like the name or an identification number, but also variables, which contain sensitive information on an individual. To protect the survey units against disclosure it might not suffice to delete the variables, which are directly linked to entities, because some of the units might still be identifiable by the rest of their records. Statistical disclosure control is nothing else but a balancing act between the protection of the anonymity of the survey units and the preservation of information contained in the data (cf. Skinner, Marsh, Openshaw and Wymer 1994). Methods of data masking can be classified into three categories (cf. Domingo-Ferrer and Mateo-Sanz 2002 or Winkler 2004): (1) The *global recoding* of variables into less detailed categories or larger intervals (see for instance: Willenborg and de Waal 1996, page 5f) or the *local recoding* using different grouping schemes at unit level (cf. Hua and Pei 2008, page 215f). (2)

The *local suppression* of certain variables for survey units with a high risk of re-identification by simply setting their values at "missing" (cf. Willenborg and de Waal 1996, page 77). (3) The *substitution* of true values of a variable by other values.

One of the strategies of the third category is the *micro-aggregation* of variables (cf. Defays and Anwar 1998). Therein the true variable values are for example sorted by size and then divided into (small) groups. Within each group data aggregates are released instead of the original observations. Another such method is *data-swapping*, where data from units with a high risk of re-identification are interchanged with data from another subset of survey units (cf. Dalenius and Reiss 1982). Another technique of substituting identifying or sensitive information is the *addition of noise* to the observed values, meaning that the outcome of a random experiment is added to each datum (cf. Dalenius 1977 or Fuller 1993). Finally also the randomized response techniques can be used to mask identifying or sensitive variables. In this case either the survey units already perform the data masking at the survey's design stage or the statistical agency applies the probability mechanism of the technique before the release of the microdata file (cf. Rosenberg 1980, Kim 1987, Gouweleuw, Kooiman, Willenborg and de Wolf 1998, or van den Hout and van der Heijden 2002).

All methods of statistical disclosure control protect the survey units' privacy by a loss of information, which can be seen as the price that has to be paid for it. To be able to appropriately adjust the estimation process the user of the microdata file has to be informed about the details of the masking procedure.

A new standardization of the techniques of randomized response follows in Section 2 of this paper. Furthermore the statistical properties of the standardized estimator are derived for general probability sampling. In Section 3 the essential perspective of privacy protection is described. The question, which of the special cases included in the standardization is most efficient, is answered in the subsequent Section 4. Section 5 contains a real-data example, which demonstrates the application of the recommendations of Section 4 in a survey on academic cheating behaviour.

2. Standardizing randomized response strategies

Let us formulate the following standardization of the randomized response strategies: Each respondent has either to answer randomly with probability

- p_1 the question "Are you a member of group U_A ?",
 - p_2 the question "Are you a member of group U_A^c ?"
- or

- p_3 the question “Are you a member of group U_B ?” or is instructed just to say
 - “yes” with probability p_4 or
 - “no” with probability p_5

($\sum_{i=1}^5 p_i = 1$, $0 \leq p_i \leq 1$ for $i = 1, 2, \dots, 5$). The N_B elements of group U_B are characterized by the possession of a completely innocuous attribute B (for instance a season B of birth), that should not be related to the possession or nonpossession of attribute A . This nonsensitive question on membership of group U_B was introduced as an alternative to the question on membership of U_A by Horvitz, Shah and Simmons (1967) to further reduce the respondent's perception of the sensitivity of the procedure. $\pi_B = N_B/N$ (with $0 < \pi_B < 1$) is the relative size of group $U_B \cdot \pi_B$ and the probabilities p_1, p_2, \dots, p_5 are the *design parameters* of our standardized randomized response technique.

Let

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ answers “yes”,} \\ 0 & \text{otherwise} \end{cases}$$

($i = 1, 2, \dots, n$). For an element i the probability of a “yes”-answer with respect to the randomized response questioning design R is for given x :

$$P_R(y_i = 1) = p_1 \cdot x_i + p_2 \cdot (1 - x_i) + p_3 \cdot \pi_B + p_4 = a \cdot x_i + b \quad (4)$$

with $a \equiv p_1 - p_2$ and $b \equiv p_2 + p_3 \cdot \pi_B + p_4$. Then the term

$$\hat{x}_i = \frac{y_i - b}{a}$$

is unbiased for the true value x_i ($a \neq 0$). Using these “substitutes” for x_i (and assuming full cooperation of the respondents) the following theorems apply:

Theorem 1: For a probability sampling design with inclusion probabilities π_i the following unbiased estimator of parameter π_A is given:

$$\hat{\pi}_A = \frac{1}{N} \cdot \sum_s \frac{\hat{x}_i}{\pi_i} \quad (5)$$

Theorem 2: For a probability sampling design P the variance of the standardized estimator $\hat{\pi}_A$ (5) is given by

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left(V_P \left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b \cdot (1 - b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i} \right) \quad (6)$$

For the proofs of both theorems see the Appendix. The first summand within the outer brackets of (6) refers to the

variance of the Horvitz-Thompson estimator for the total $\sum_U x_i$ for a probability sampling design P when the question on membership of U_A is asked directly. The second one can be seen as the price we have to pay in terms of accuracy for the privacy protection offered by the randomized response questioning design. Apparently this variance can be estimated unbiasedly by inserting an unbiased estimator $\hat{V}_P(\sum_s x_i / \pi_i)$ for $V_P(\sum_s x_i / \pi_i)$ and $\sum_s \hat{x}_i / \pi_i^2$ for $\sum_U x_i / \pi_i$.

For simple random sampling without replacement for instance estimator (5) is given by

$$\hat{\pi}_A = \frac{\hat{\pi}_y - b}{a} \quad (7)$$

with $\hat{\pi}_y = \sum y_i / n$, the proportion of “yes”-answers in the sample. In this case the variance (6) of the standardized estimator $\hat{\pi}_A$ is given by

$$V(\hat{\pi}_A) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{1}{n} \cdot \left(\frac{b \cdot (1 - b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \pi_A \right) \quad (8)$$

This theoretical variance is unbiasedly estimated by

$$\hat{V}(\hat{\pi}_A) = \frac{\hat{\pi}_A \cdot (1 - \hat{\pi}_A)}{n - 1} \cdot \frac{N - n}{N} + \frac{1}{n} \cdot \left(\frac{b \cdot (1 - b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \hat{\pi}_A \right) \quad (9)$$

To be able to calculate $\hat{\pi}_A$ at all, the question on membership of U_A (or U_A^c , but we will ignore this possibility subsequently without loss of generality) must be included in the questioning design with $p_1 > 0$. There is a total of 16 combinations of this question with the four other questions or answers (see: Table 1). These combinations can be described as special cases of our standardized response strategy. For example choosing $p_1 = 1$ leads to the direct questioning on the subject. If we let $0 < p_1 < 1$ and $p_2 = 1 - p_1$ the standardized questioning design turns into Warner's procedure. For $0 < p_1 < 1$ and $p_3 = 1 - p_1$ one gets Horvitz *et al.*'s technique with known π_B (see: Greenberg, Abul-El, Simmons and Horvitz 1969). (For other special cases already published as to the best of our knowledge, the reader is referred to the “References”-column of Table 1).

The question, that arises directly from these considerations, is how to choose the design parameters of the standardized response technique to find out the strategies that perform best. We will answer this question in Section 4. But for this purpose we have to include the level of privacy protection, which results from choosing these parameters differently, in our considerations.

Table 1

All special cases of the standardized randomized response strategy

Design	Questions/Answers					References
	U_A	U_{A^c}	U_B	yes	no	
ST1	•					Direct questioning
ST2	•	•				Warner (1965) ¹
ST3			•			Greenberg <i>et al.</i> (1969) ²
ST4	•			•		
ST5	•				•	
ST6	•	•	•			
ST7	•	•		•		
ST8	•	•				• Quatember (2007) ³
ST9	•		•	•		
ST10	•		•			• Singh, Horn, Singh and Mangat (2003) ⁴
ST11	•		•	•		• Fidler and Kleinknecht (1977) ⁵
ST12	•	•	•	•		
ST13	•	•	•		•	
ST14	•	•		•	•	
ST15	•		•	•	•	
ST16	•	•	•	•	•	

1. A two-stage version was presented by Mangat and Singh (1990)
2. A two-stage version was presented by Mangat (1992)
3. This is a one-stage version of Mangat, Singh and Singh (1993)
4. This is a one-stage version of Singh, Singh, Mangat and Tracy (1994)
5. A two-stage version was presented by Singh, Singh, Mangat and Tracy (1995)

3. Privacy protection

To be able to compare the efficiency of questioning designs with different design parameters it is apparently inevitable to measure the loss of the respondents' privacy induced by these parameters. The following ratios λ_1 and λ_0 of conditional probabilities may be used for this purpose (*cf.* for example the similar "measures of jeopardy" in Leysieffer and Warner 1976, page 650):

$$\lambda_j = \frac{\max[P(y_i = j \mid i \in U_A), P(y_i = j \mid i \in U_A^c)]}{\min[P(y_i = j \mid i \in U_A), P(y_i = j \mid i \in U_A^c)]} \quad (10)$$

($1 \leq \lambda_j \leq \infty$; $j = 1, 0$).

For $j = 1$ (10) refers to the privacy protection with respect to a "yes", for $j = 0$ with respect to a "no"-answer. For the standardized questioning design these " λ -measures" of loss of privacy are given by

$$\lambda_1 = \frac{\max[a + b; b]}{\min[a + b; b]} \quad (11)$$

and

$$\lambda_0 = \frac{\max[1 - (a + b); 1 - b]}{\min[1 - (a + b); 1 - b]} \quad (12)$$

$\lambda_1 = \lambda_0 = 1$ indicates a totally protected privacy. This means that the answer of the responding unit contains absolutely no information on the subject under study. This applies for $a = 0$. The more the λ -measures differ from

unity, the more information about the characteristic under study is contained in the answer on the record. At the same time the efficiency of the estimation increases (see below), but the individual's protection against the data collector decreases. For the direct questioning design with $p_1 = 1$, where no masking of the variable is done at all, these measures are given by $\lambda_1 = \lambda_0 = \infty$.

Let the values $\lambda_{1, \text{opt}}$ and $\lambda_{0, \text{opt}}$ be the maximum λ -values of (11) and (12), that the agency considers to allow enough disclosure protection for the records. In the case of the strategy's usage as to avoid nonresponse and untruthful answering in surveys we may also model the respondents' willingness to cooperate as a function of perceived privacy protection. If the privacy of the respondents is sufficiently protected by the randomization device their full cooperation is assumed. Exceeding the limits $\lambda_{1, \text{opt}}$ and/or $\lambda_{0, \text{opt}}$ would then automatically introduce untruthful answering and nonresponse into the survey and therefore set us back to the starting point of the problem. Fidler and Kleinknecht (1977) showed in their study for design ST11 (Table 1) containing nine variables of very different levels of sensitivity, that their choice of the design parameters ($p_1 = 10/16$, $p_4 = p_5 = 3/16$) yielded nearly full and truthful response for each variable including sexual behaviour (*ibid.*, page 1048). Inserting these values in (11) and (12) gives $\lambda_1 = \lambda_0 = 13/3$. This finding corresponds in the main with results that can be derived from the experiment by Soeken and Macready (1982) and with recommendations given by Greenberg *et al.* (1969). Therefore choosing $\lambda_{1, \text{opt}}$ and/or $\lambda_{0, \text{opt}}$ close to a value of 4 could be a good choice for most variables, when the standardized randomized response method is used to avoid refusals and untruthful answering of respondents in a survey.

Without loss of generality let us assume subsequently, that we will choose the two categories of the variable under study in such way, that the membership of U_A is at least as sensitive as the membership of U_A^c ($1 \leq \lambda_{1, \text{opt}} \leq \lambda_{0, \text{opt}} \leq \infty$). From (11) and (12) the terms a and b can be expressed by the λ -values λ_1 and λ_0 . Their sum is given by:

$$a + b = \frac{1 - \frac{1}{\lambda_0}}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}} \quad (13)$$

with

$$b = \frac{\frac{1}{\lambda_1} \cdot \left(1 - \frac{1}{\lambda_0}\right)}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}} \quad (14)$$

and

$$a = \frac{\left(1 - \frac{1}{\lambda_1}\right) \cdot \left(1 - \frac{1}{\lambda_0}\right)}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}}. \quad (15)$$

We keep the double ratios on the right of (14) and (15) to find easily the limits for $\lambda_1 \rightarrow \infty$ and $\lambda_0 \rightarrow \infty$ respectively.

This means that for a given sampling design P the extent of the term $(b \cdot (1 - b) / a^2) \cdot \sum_U (1 / \pi_i) + (1 - 2 \cdot b - a / a) \cdot \sum_U (x_i / \pi_i)$ in the variance expression (6) does not depend on a single value of the design parameters, but on their aggregated effect on the loss of privacy measured by λ_1 and λ_0 . Questioning designs with the same λ -values are equally efficient. Designs with larger λ_1 and/or λ_0 are less efficient than designs with lower λ 's.

4. Optimum questioning designs

It does depend on the type of re-identification risk or sensitivity of the subject under study which of the special cases of the standardized randomized response strategy of Table 1 can be most efficient for given λ -measures. Strategies *ST5* and *ST8* can never perform best, because they do always protect a "no"-answer more than a "yes".

For a nonidentifying (or nonsensitive) variable (like for instance the season of birth), where $\lambda_{1, \text{opt}} = \lambda_{0, \text{opt}} = \infty$ applies, only the direct questioning design (*ST1* of Table 1) can achieve the variance-optimum performance (see Table 2, which shows these values of the design parameters, which guarantee the best performance of the estimator $\hat{\pi}_A$; to be able to use Table 2 properly the categorical variable under study has to be classified according to the following categories: C_1 : The variable is not sensitive at all ($\lambda_{1, \text{opt}} = \lambda_{0, \text{opt}} = \infty$); C_2 : Only the membership of group U_A is sensitive, but not of U_A^c ($\lambda_{1, \text{opt}} < \lambda_{0, \text{opt}} = \infty$); C_3 : The membership of both groups U_A and U_A^c is sensitive, but not equally ($\lambda_{1, \text{opt}} < \lambda_{0, \text{opt}} < \infty$); C_4 : The membership of U_A and of U_A^c is equally sensitive ($\lambda_{1, \text{opt}} = \lambda_{0, \text{opt}} < \infty$), which shows these values of the design parameters, which guarantee the best performance of the estimator $\hat{\pi}_A$). Although the other designs can be used for such variables, they do unnecessarily protect the privacy of the respondents in some way. This has to be paid by a loss of accuracy of the estimation of π_A . But for $p_1 = 1$ ($a = 1$ and $b = 0$) the variance of $\hat{\pi}_A$ (5) turns to the common formula of the direct questioning with the assumption of full response: $V_P(\hat{\pi}_A) = 1/N^2 \cdot V_P(\sum_s x_i / \pi_i)$.

For a variable, of which only the membership of U_A , but not of U_A^c is sensitive (for instance: U_A = set of drug users within the last year; $U_A^c = U - U_A$) there is $\lambda_{1, \text{opt}} < \lambda_{0, \text{opt}} = \infty$. Calculating (14) and (15) for $1 < \lambda_1 < \infty$ and $\lambda_0 \rightarrow \infty$ gives $a = 1 - b$ and inserting this into (6) leads to the following expression for the variance of the estimator:

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left[V_P \left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b}{1-b} \cdot \left(\sum_U \frac{1}{\pi_i} - \sum_U \frac{x_i}{\pi_i} \right) \right]. \quad (16)$$

Looking for those values of the design parameters, for which the standardized randomized response strategy can achieve this variance and for which equations (14) to (15) hold, we do find that in this case there is only one solution! The only questioning design, that is able to perform optimally, is *ST4*. Its variance-optimum design parameters are given by $p_1 = (\lambda_1 - 1) / \lambda_1$ and $p_4 = 1 - p_1$ (see Table 2). This means, that with probability $p_1 = (\lambda_1 - 1) / \lambda_1$ a respondent is asked the question on membership of U_A and with the remaining probability he or she is instructed to say "yes". In this way the data collector is only able to conclude from a "no"-answer directly on the nonsensitive non-possession of A but not from a "yes"-answer on the possession of this sensitive or identifying attribute.

Questioning design *ST1* is not applicable for such subjects, because it does not protect the respondent's privacy in case of a "yes"-answer at all. All the other procedures protect a "no"-answer more than necessary. Therefore they may be used, but they cannot achieve the efficiency of *ST4*.

If the membership of both U_A and U_A^c is sensitive, so that the variable is sensitive as a whole (for instance: U_A = set of married people, who had at least one sexual intercourse with their partners last week; $U_A^c = U - U_A$), $\lambda_{1, \text{opt}} \leq \lambda_{0, \text{opt}} < \infty$ applies. In this case neither the direct questioning on the subject nor design *ST4* can be used because they are not able to protect both possible answers.

The other designs are applicable for such topics, but Warner's design cannot achieve the efficiency of the others, if $\lambda_{1, \text{opt}} < \lambda_{0, \text{opt}}$. The reason is that this design always protects the respondent's privacy with respect to a "yes"-answer equally to a "no"-answer. But if $\lambda_{1, \text{opt}} = \lambda_{0, \text{opt}}$ despite to the claims of some publications in the past (see for instance: Greenberg *et al.* 1969, page 526f, Mangat and Singh 1990, page 440, Singh *et al.* 2003, page 518f) there is *not one* randomized response technique that can perform *better* than Warner's technique *ST2* with the optimum design parameters p_1 and p_2 according to Table 2. For *ST7* this is only valid for $\lambda_{1, \text{opt}} < \lambda_{0, \text{opt}}$. Therefore *ST7* is the perfect supplement of *ST2*, for which the very opposite is true.

Table 2
Optimum design parameters for given λ_1 and λ_0 and different types of sensitivity of the variable under study

Questioning design (Subject category)	Variance-optimum design parameters
ST1 (C_1)	$p_1 = 1$
ST2 (C_4)	$p_1 = \frac{\lambda_1}{\lambda_1 + 1}, p_2 = 1 - p_1$
ST3 (C_3, C_4)	$\pi_B = \frac{\lambda_0 - 1}{\lambda_1 + \lambda_0 - 2}, p_1 = \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1}, p_3 = 1 - p_1$
ST4 (C_2)	$p_1 = \frac{\lambda_1 - 1}{\lambda_1}, p_4 = 1 - p_1$
ST6 (C_4)	$\pi_B = 0.5, p_1: \frac{\lambda_1 - 1}{\lambda_1 + 1} < p_1 < \frac{\lambda_1}{\lambda_1 + 1}, p_2 = p_1 - \frac{\lambda_1 - 1}{\lambda_1 + 1},$ $p_3 = 1 - p_1 - p_2$
ST6 (C_3)	$\pi_B: \frac{\lambda_0 - 1}{\lambda_1 + \lambda_0 - 2} < \pi_B < 1,$ $p_1 = \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1} + \frac{(\lambda_1 - 1)\pi_B - (\lambda_0 - 1)(1 - \pi_B)}{(\lambda_1 \cdot \lambda_0 - 1)(2\pi_B - 1)},$ $p_2 = p_1 - \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1}, p_3 = 1 - p_1 - p_2$
ST7 (C_3)	$p_1 = \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1}, p_2 = \frac{\lambda_1 - 1}{\lambda_1 \cdot \lambda_0 - 1}, p_4 = 1 - p_1 - p_2$
ST9 (C_3, C_4)	$\pi_B: 0 < \pi_B < \frac{\lambda_0 - 1}{\lambda_1 + \lambda_0 - 2}, p_1 = \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1},$ $p_3 = \frac{\lambda_1 - 1}{(\lambda_1 \cdot \lambda_0 - 1)(1 - \pi_B)}, p_4 = 1 - p_1 - p_3$
ST10 (C_3, C_4)	$\pi_B: \frac{\lambda_0 - 1}{\lambda_1 + \lambda_0 - 2} < \pi_B < 1, p_1 = \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1},$ $p_3 = \frac{\lambda_0 - 1}{(\lambda_1 \cdot \lambda_0 - 1)\pi_B}, p_5 = 1 - p_1 - p_3$
ST11 (C_3, C_4)	$p_1 = \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1}, p_4 = \frac{\lambda_0 - 1}{\lambda_1 \cdot \lambda_0 - 1}, p_5 = 1 - p_1 - p_4$
ST12 (C_3, C_4)	$p_1: \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1} < p_1 < \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1}, p_2 = p_1 - \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1},$ $\pi_B: 0 < \pi_B < \frac{\lambda_0 - 1 - p_2(\lambda_1 \cdot \lambda_0 - 1)}{\lambda_1 + \lambda_0 - 2 - 2p_2(\lambda_1 \cdot \lambda_0 - 1)}, p_3 = \frac{\lambda_1 - 1 - p_2(\lambda_1 \cdot \lambda_0 - 1)}{(\lambda_1 \cdot \lambda_0 - 1)(1 - \pi_B)},$ $p_4 = 1 - \sum_{i=1}^3 p_i$
ST13 (C_3, C_4)	$p_1: \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1} < p_1 < \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1}, p_2 = p_1 - \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1},$ $\pi_B: \frac{\lambda_0 - 1 - p_2(\lambda_1 \cdot \lambda_0 - 1)}{\lambda_1 + \lambda_0 - 2 - 2p_2(\lambda_1 \cdot \lambda_0 - 1)} < \pi_B < 1, p_3 = \frac{\lambda_0 - 1 - p_2(\lambda_1 \cdot \lambda_0 - 1)}{(\lambda_1 \cdot \lambda_0 - 1)\pi_B},$ $p_5 = 1 - \sum_{i=1}^3 p_i$
ST14 (C_3, C_4)	$p_1: \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1} < p_1 < \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1}, p_2 = p_1 - \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1},$ $p_4 = \frac{\lambda_0 - 1}{\lambda_1 \cdot \lambda_0 - 1} - p_2, p_5 = 1 - p_1 - p_2 - p_4$
ST15 (C_3, C_4)	$\pi_B: 0 < \pi_B < 1, p_1 = \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1},$ $p_3: 0 < p_3 < \frac{\lambda_1 - 1}{(\lambda_1 \cdot \lambda_0 - 1)(1 - \pi_B)}, p_4 = \frac{\lambda_0 - 1}{\lambda_1 \cdot \lambda_0 - 1} - p_3 \cdot \pi_B,$ $p_5 = 1 - p_1 - p_3 - p_4$
ST16 (C_3, C_4)	$\pi_B: 0 < \pi_B < 1, p_1: \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1} < p_1 < \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1},$ $p_2 = p_1 - \frac{(\lambda_1 - 1)(\lambda_0 - 1)}{\lambda_1 \cdot \lambda_0 - 1}, p_3: 0 < p_3 < \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1} - p_1,$ $p_4 = \frac{\lambda_0 - 1}{\lambda_1 \cdot \lambda_0 - 1} - p_2 - p_3 \cdot \pi_B, p_5 = 1 - \sum_{i=1}^4 p_i$

All others of the designs of Table 1 like $ST11$ or $ST14$ can perform equally efficient for $\lambda_{1,\text{opt}} \leq \lambda_{0,\text{opt}} < \infty$, if the design parameters are chosen according to the restrictions (14) to (15). Among them Greenberg *et al.*'s strategy with known π_B ($ST3$) has on the one hand the advantage over Warner's design to be able to perform optimally also if $\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}}$. On the other hand, however, it has the disadvantage (like $ST6$), that the size π_B of subpopulation U_B is completely predetermined (or at least bounded by an interval), if we want to achieve the optimum efficiency. This means in practice, that we have to find a subpopulation not related to the possession and nonpossession of attribute A and of appropriate relative size to be able to achieve the estimator's optimum accuracy. In principle this also applies to $ST9$, $ST10$, $ST12$ and $ST13$, but looking at the presettings of design parameter π_B , it turns out that $ST9$ and $ST10$ as well as $ST12$ and $ST13$ perfectly complement each other so that in fact any subset U_B of the population can be used. Finally the most complex special cases, $ST15$ and $ST16$, of our standardized randomized response strategy can both be used with any subpopulation $U_B \subset U$ to achieve the best performance.

5. A real-data example

An empirical study was carried out to demonstrate the applicability of the strategy as a questioning design. For this purpose the population of 80 students, who attended the author's course on "Statistics II" at the Johannes Kepler University in Linz (Austria) during the spring term of 2009, volunteered for the survey. The subject under study was academic cheating behaviour. To this end cheating was defined as any behaviour, that was not allowed in the written exams (including just looking at the test scripts of other students or the use of forbidden documents). It is beyond doubt that this subject is sensitive for such a population. Moreover during the survey all of the students were sitting in one lecture room. The parameter of interest was the proportion of the population of students, that fudged on at least one of the exams of the previous semester (including the exam of the author's course on "Statistics I"). Therefore it is beyond reasonable doubt to assume, that direct questioning on the subject would have resulted into a substantial underestimation of this proportion. An empirical study of Scheers and Dayton (1987) for instance showed very small proportions for almost all different cheating behaviours asked, when the subject in question was asked directly. The use of Greenberg's randomized response strategy $ST3$ lead to a significant increase of these proportions (ibid., page 68).

Apparently, for the variable of interest the membership of group U_A , formed by the "cheaters", is sensitive, but not

the membership of the complementary set U_A^c . Therefore in accordance with the recommendations of Section 4 we decided to use questioning design $ST4$ for our survey and to compare it with Warner's strategy $ST2$. The λ -values of loss of privacy were fixed at $\lambda_1 = 4$ and $\lambda_0 = \infty$. From Table 2 we calculated $p_1 = 0.75$ and $p_4 = 0.25$ as the variance-optimum design parameters of $ST4$. To achieve these probabilities the students were asked to throw two dice without showing the result to somebody else and answer in a questionnaire the question "Did you cheat at the exams at least one time?" only if the sum of the numbers on the dice was 5 to 10. Otherwise they should just respond "yes".

Previous to the survey some effort was made to explain the consequences of this randomization strategy on the privacy protection. After giving the answer on the first sheet of the questionnaire, only these sheets were collected. 63 out of the 80 persons answered "yes". 20 of 80 students were expected to do so, because they received the "say yes-instruction". Therefore expected 43 of 60 other students should have answered "yes" on the sensitive question. The estimator for π_A is given by

$$\hat{\pi}_A^{ST4} = \frac{\hat{\pi}_y^{ST4} - p_4}{p_1} = \frac{0.7875 - 0.25}{0.75} = 0.716.$$

For this population survey the estimated variance of $\hat{\pi}_A$ is then

$$\widehat{V}(\hat{\pi}_A^{ST4}) = \frac{1 - p_1}{n \cdot p_1} \cdot (1 - \hat{\pi}_A^{ST4}) = 1.181 \cdot 10^{-3}.$$

After this questioning design was completed, the students were asked directly on the second sheet of the questionnaire, whether they had truthfully answered the first question or not. Only four students said that this was not the case. This means, that – if that's true – it is likely that 4 more students did actually cheat. The next question to answer was, if they would still cooperate, if p_1 (of $ST4$) would be higher than 0.75. 32 of 80 students agreed to do so, but the others did not. Obviously (at least) four of them did not cooperate when p_1 was 0.75.

Finally, Warner's technique was applied with the same sensitive question as $ST4$ before. To come close to a λ_1 -level of 4 – indicating the same loss of privacy as to a "yes"-answer for both questioning designs –, the sum of the numbers of two dice had to be 3 to 9 to apply a design parameter $p_1 = 0.805$. The λ -measures of loss of privacy for this choice are given by $\lambda_1 = \lambda_0 = 4.143$, indicating a slightly higher loss of privacy compared to $ST4$. With a probability of 0.805 the students had to answer "Are you a member of U_A ?" and with the remaining probability the alternative "Are you a member of U_A^c ?".

Now only 38 of 80 persons gave a "yes"-answer. This results in an estimated proportion of "cheaters" of

$$\hat{\pi}_A^{ST2} = \frac{\hat{\pi}_y^{ST2} - p_2}{p_1 - p_2} = \frac{0.475 - 0.194}{0.61} = 0.4590.$$

Additionally to the slight increase of the objective loss of privacy there is another reasonable explanation for this significantly lower result. Although λ_1 did not change that much, some test persons must have been irritated by the raise of p_1 up to 0.805 after being asked for $ST4$, if they would still cooperate, if p_1 would be higher than 0.75. Not being able to distinguish between the loss of privacy caused by different design parameters in different questioning designs, some of the “cheaters” did not want to answer truthfully again. Just to demonstrate the effect of the different questioning designs on the efficiency of the estimation process we calculate the estimator of the variance of $\hat{\pi}_A^{ST2}$:

$$\hat{V}(\hat{\pi}_A^{ST2}) = \frac{p_1 \cdot (1 - p_1)}{n \cdot (2p_1 - 1)^2} = 5.243 \cdot 10^{-3}.$$

The reason for this considerable increase of the estimated variance is, that Warner's strategy does protect a “no”-answer always in the same way as a “yes”. Since in our case a “no”-answer does not have to be protected at all, this unnecessary protection has to be paid in terms of accuracy.

6. Summary

Randomized response strategies have originally been developed to reduce the nonresponse as well as the untruthful answering rate for sensitive subjects in sample surveys, but they can be applied as masking techniques for public use microdata files as well. The standardization of these techniques for the estimation of proportions developed in this paper provides an opportunity to derive a general formula for the variance of the estimator under probability sampling. Different questioning designs, partly published, partly – to the best of our knowledge – unpublished up to now, can be regarded as special cases of the standardized strategy (see Table 1). For the purpose of a comparison of the accuracy of these designs it is essential to include the levels of privacy protection offered by them in our considerations. Doing this by means of the “ λ -measures” of loss of privacy explicated in Section 3 a completely new picture has to be painted in comparison to almost all publications in the past as far as the author knows them. It turns out that the identifying or sensitive subjects have to be classified into different categories in order to find the variance-minimum questioning designs for a given privacy protection (see Table 2). The first category consists of

subjects, which are not sensitive at all. The second comprises topics, where only the possession but not the nonpossession of a certain attribute is embarrassing to the respondents. The last category is formed by subjects, which are sensitive as a whole.

For subjects out of the first category it is clear enough that no strategy can be more efficient than the direct questioning on the subject ($ST1$ of Table 1).

Concerning topics of the second category there is just one design available, that can achieve the minimum variance of the estimator. This is the questioning design in which each respondent either with probability p_1 has to answer the question on membership of the sensitive group or with probability $1 - p_1$ is instructed to answer “yes” ($ST4$). All the other special cases of the standardized strategy protect the interviewee's privacy not only in case of a “yes”-answer like $ST4$ does, but also in case of a “no”-answer. Therefore their performances cannot reach the minimum achievable level.

For subjects out of the third category it is shown, that contrary to the claim of other publications, there is not one single strategy available that can perform better than Warner's of 1965 as long as the membership of the subgroup under investigation is equally sensitive to the membership of its complement. A lot of other designs are equally efficient as Warner's but not a single one is more efficient.

For the variables of this category, where the membership of one group is sensitive, but not equally sensitive as the membership of the complementary one, the situation changes dramatically: Compared under the same levels of privacy protection Warner's technique is not able to achieve the best achievable performance of the standardized randomized design anymore, whereas many other strategies can. For some of the designs including the question on membership of a nonsensitive subpopulation not related to the attribute under study, it is required to find an adequate subpopulation of predetermined relative size. Other designs can be used with subpopulations of any size and are therefore more practicable. Therefore a data collector or publisher could select that one of the equally efficient designs, that seems to be more easily applicable than the others.

Acknowledgements

The author is very grateful to the Associate Editor and two referees for their valuable comments and suggestions.

Appendix

Proofs of theorems 1 and 2

Proof of Theorem 1:

$$\begin{aligned} E(\hat{\pi}_A) &= \frac{1}{N} \cdot E_P \left(E_R \left(\sum_s \frac{\hat{x}_i}{\pi_i} \mid s \right) \right) \\ &= \frac{1}{N} \cdot E_P \left(\sum_s \frac{x_i}{\pi_i} \right) = \frac{1}{N} \cdot \sum_U x_i = \pi_A. \end{aligned}$$

The variance of estimator (5) is given by

$$V(\hat{\pi}_A) = V_P(E_R(\hat{\pi}_A \mid s)) + E_P(V_R(\hat{\pi}_A \mid s)).$$

Then

$$V_P(E_R(\hat{\pi}_A \mid s)) = \frac{1}{N^2} \cdot V_P \left(\sum_s \frac{x_i}{\pi_i} \right).$$

Let the sample inclusion indicator

$$I_i = \begin{cases} 1 & \text{if unit } i \in s, \\ 0 & \text{otherwise.} \end{cases}$$

Because the covariance $C_R(\hat{x}_i, \hat{x}_j \mid s) = 0 \ \forall \ i \neq j$, for the second summand of $V(\hat{\pi}_A)$ applies

$$\begin{aligned} E_P(V_R(\hat{\pi}_A \mid s)) &= E_P \left(\frac{1}{N^2} \cdot V_R \left(\sum_U I_i \cdot \frac{\hat{x}_i}{\pi_i} \mid s \right) \right) \\ &= E_P \left(\frac{1}{N^2} \cdot \sum_U \frac{I_i^2}{\pi_i^2} \cdot V_R(\hat{x}_i) \right) \\ &= \frac{1}{N^2} \cdot \sum_U \frac{V_R(\hat{x}_i)}{\pi_i}. \end{aligned}$$

For $V_R(\hat{x}_i)$ we have

$$V_R(\hat{x}_i) = \frac{1}{a^2} \cdot V_R(y_i)$$

and

$$\begin{aligned} V_R(y_i) &= b + a \cdot x_i - (b + a \cdot x_i)^2 \\ &= (b + a \cdot x_i) \cdot (1 - b - a \cdot x_i) \\ &= b \cdot (1 - b) + a \cdot (1 - 2 \cdot b - a) \cdot x_i. \end{aligned}$$

Then

$$\begin{aligned} E_P(V_R(\hat{\pi}_A \mid s)) &= \\ \frac{1}{N^2} \cdot \left(\frac{b \cdot (1 - b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i} \right). \end{aligned}$$

This completes the proof of Theorem 2.

References

- Chaudhuri, A., and Mukerjee, R. (1987). *Randomized Response*. New York: Marcel Dekker.
- Dalenius, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86.
- Dalenius, T., and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.
- Defays, D., and Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14 (4), 449-461.
- Domingo-Ferrer, J., and Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.
- Fidler, D.S., and Kleinknecht, R.E. (1977). Randomized response versus direct questioning: Two data collection methods for sensitive information. *Psychological Bulletin*, 84 (5), 1045-1049.
- Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9 (2), 383-406.
- Gouweleew, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14 (4), 463-478.
- Greenberg, B.G., Abul-El, A.-L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken: John Wiley & Sons, Inc.
- Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 65-72.
- Hua, M., and Pei, J. (2008). A survey of utility-based privacy-preserving data transformation methods. In: *Privacy-preserving Data Mining: Models and Algorithms*, (Eds., C.C. Aggarwal and P.S. Yu), New York: Springer, 207-238.
- Kim, J. (1987). A further development of the randomized response technique for masking dichotomous variables. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 239-244.
- Kim, J.M., and Warde, W.D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, 133, 211-221.
- Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77 (2), 436-438.
- Landsheer, J.A., van der Heijden, P. and van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, 33, 1-12.

- Leysieffer, F.W., and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- Mangat, N.S. (1992). Two stage randomized response sampling procedure using unrelated question. *Journal of the Indian Society of Agricultural Statistics*, 44, 82-87.
- Mangat, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society, Series B*, 56, 93-95.
- Mangat, N.S., and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.
- Mangat, N.S., Singh, S. and Singh, R. (1993). On the use of a modified randomization device in randomized response inquiries. *Metron*, 51, 211-216.
- Nathan, G. (1988). A bibliography of randomized response: 1965-1987. *Survey Methodology*, 14, 331-346.
- Quatember, A. (2007). Comparing the efficiency of randomized response techniques under uniform conditions. *IFAS Research Paper Series*, 23, www.ifas.jku.at/e2550/e2756/index_ger.html.
- Rosenberg, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 311-316.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.
- Scheers, N.J., and Dayton, C.M. (1987). Improved estimation of academic cheating behaviour using the randomized response technique. *Research in Higher Education*, 26 (1), 61-69.
- Singer, E., Mathiowetz, N.A. and Couper, M.P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Census. *The Public Opinion Quarterly*, 57 (4), 465-482.
- Singer, E., van Hoewyk, J. and Neugebauer, R.J. (2003). Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 Census. *The Public Opinion Quarterly*, 67 (3), 368-384.
- Singh, R., Singh, S., Mangat, N.S. and Tracy, D.S. (1995). An improved two stage randomized response strategy. *Statistical Papers*, 36, 265-271.
- Singh, S., Horn, S., Singh, R. and Mangat, N.S. (2003). On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in Transition*, 6 (4), 515-522.
- Singh, S., Singh, R., Mangat, N.S. and Tracy, D.S. (1994). An alternative device for randomized responses. *Statistica*, 54, 233-243.
- Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10 (1), 31-51.
- Soeken, K.L., and Macready, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92 (2), 487-489.
- Tracy, D.S., and Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade – A follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4 (2/3), 147-158.
- van den Hout, A., and van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, 70 (2), 269-288.
- Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- Willenborg, L., and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York: Springer.
- Winkler, W.E. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. *Research Report Series of the Statistical Research Division of the U.S. Bureau of the Census*, #2004-06.

Treatments for link nonresponse in indirect sampling

Xiaojian Xu and Pierre Lavallée¹

Abstract

We examine overcoming the overestimation in using generalized weight share method (GWSM) caused by link nonresponse in indirect sampling. A few adjustment methods incorporating link nonresponse in using GWSM have been constructed for situations both with and without the availability of auxiliary variables. A simulation study on a longitudinal survey is presented using some of the adjustment methods we recommend. The simulation results show that these adjusted GWSMs perform well in reducing both estimation bias and variance. The advancement in bias reduction is significant.

Key Words: Weight share method; Nonresponse; Indirect sampling; Longitudinal survey.

1. Introduction

Indirect sampling refers to selecting samples from the population which is not, but it is related to, the target population of interest. Such a sampling scheme is often carried out when we do not have sampling frames for the target population, but have sampling frames for another population which is related to it. We call the latter sampling population. For an example in Lavallée (2007), we consider the situation where the estimate is concerned with young children belonging to families, but we only have a list of parents' names as our sampling frame. Consequently, we must first select a sample of parents before we can select the sample of children. In this typical indirect sampling situation. The sampling population is that of parents while the target population is that of children. We note that the children of a particular family can be selected through either the father or the mother. Figure 1 provides a simple illustration for this indirect sampling scheme (Figure 1.2, Lavallée 2007).

There is a sizeable amount of literature concerning estimation problems that are associated with indirect sampling, a few of which we name here. Initially, estimation methods for production of cross-sectional estimates using longitudinal household survey are discussed in Ernst (1989). This study presents weight share method in the context of longitudinal survey and also shows that this method provides an unbiased estimator for the total for any characteristic in the population of interest. Kalton and Brick (1995) conclude that such a method also provides minimal variance of estimated population total for some simple sampling schemes for the longitudinal household panel survey. Lavallée (1995) extends weight share method in a completely general context of indirect sampling which includes longitudinal survey as its particular example, called generalized weight share method (GWSM). This work justifies that this weighting scheme provides unbiased

estimates irrespective of sampling schemes in obtaining a sample in the sampling population. As with any other weighting scheme, in the process of GWSM implementation an adjustment for a variety of nonresponse problems must be made. Lavallée (2001) provides adjusted GWSM incorporating possible total nonresponse problems in indirect sampling. In indirect sampling there is another type of nonresponse called link nonresponse, termed by Lavallée (2001) as "relationship nonresponse," which is associated with a situation where it is impossible to determine, or where one has failed to determine, whether or not a unit in the sampling population is related to a unit in the target population. Lavallée (2001) points out the problem of overestimation in using GWSM when link nonresponse occurs and leaves finding suitable adjustment of GWSM for link nonresponse as a rather open question. This present study focuses on developing treatments of estimation bias caused by such link nonresponse.

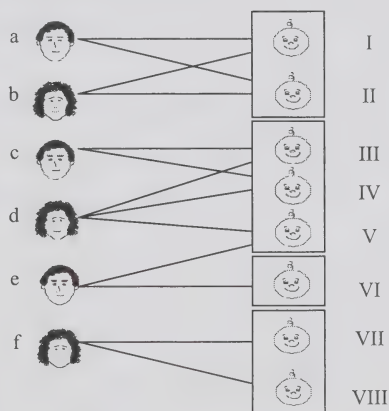


Figure 1 Indirect sampling of children

1. Xiaojian Xu, Department of Mathematics, Brock University, St. Catharines, Ontario, Canada, L2S 3A1. E-mail: xxu@brocku.ca; Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: Pierre.lavallee@statcan.gc.ca.

The rest of this work has been arranged in the following sections. Notation and the problem defined are described in Section 2. We propose a few modification methods in using GWSM incorporating link nonresponse in Section 3. A simulation study using a real life data set is presented in Section 4 with a few closing remarks in Section 5. We note that we show the advances of the new methods provided in this paper through a simulation study while other theoretical contributions relevant to this problem can be found in Lavallée (2002), Deville and Lavallée (2006), and Lavallée (2007).

2. Notation and problem

We use U^A and U^B to denote sampling population and target population respectively. Then, U^A is the population related to U^B with a known sampling frame. We let s^A, M^A , and m^A be a selected sample from U^A , the number of units in U^A , and the number of units in s^A respectively. We use π_j^A to represent the selection probability of j^{th} unit in U^A with $\pi_j^A > 0$ and $\sum_{j=1}^{M^A} \pi_j^A = m^A$. We also make use of the notation: M^B, N, U_i^B , and M_i^B to be the number of units in U^B , the number of clusters in U^B , the i^{th} cluster of U^B with $\cup_{i=1}^N U_i^B = U^B$, and the number of units in i^{th} cluster U_i^B .

We define $l_{j,ik}$ as an indicator variable of link existence: $l_{j,ik} = 1$ indicates that there is a link between j^{th} unit in U^A and k^{th} unit in U_i^B , while $l_{j,ik} = 0$ indicates otherwise. We also define $L_{j,i}^B$ as the total number of links existing between unit j of U^A and units of U_i^B , i.e., $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik}$. Let L_i^B be the total number of links existing between units of U^A and units of U_i^B , i.e., $L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B$. We denote the value of the characteristics for the k^{th} unit of i^{th} cluster in population U^B by y_{ik} , and the total of all y_{ik} s by Y^B . Then, we have $Y^B = \sum_{i=1}^N \sum_{k=1}^{M_i^B} y_{ik}$.

We let Ω^B denote the clusters in U^B where there is at least one unit ik such that $l_{j,ik} = 1$ for some j^{th} unit in s^A , and we say that it can be identified by units j in s^A , i.e., such i satisfies $L_i^B = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$. The number of clusters in Ω^B is n . After sampling we relabeled the clusters in Ω^B as $i = 1, 2, \dots, n$. We let w_{ik} refer to the estimation weight assigned to k^{th} unit of i^{th} cluster, Ω_i^A refer to the set of units in U^A that have links to some units in U_i^B with $i \in \Omega^B$, and Ω^A refer to the set of units in U^A that have links to some units in Ω^B , i.e., $\Omega^A = \{j \mid \sum_{i \in \Omega^B} L_{j,i}^B \neq 0\}$. We use s_i^A to indicate the set of units in s^A that have links to some units in U_i^B with $i \in \Omega^B$. We let T^A, T_i^A , and m_i^A denote the number of units in Ω^A , the number of units in Ω_i^A , and the number of units in s_i^A respectively. Finally, we make use of the following three indicators: let t_j be the indicator variable of being selected in s^A : $t_j = 1$ indicates

that j^{th} unit in U^A is in s^A and $t_j = 0$ indicates otherwise; let $t_{j,i}^L$ be the indicator variable of being included in s^A for units in Ω^A : $t_{j,i}^L = 1$ indicates that j^{th} unit in Ω^A is in s^A and $t_{j,i}^L = 0$ indicates otherwise; and let $t_{j,i}^L$ be the indicator variable of being included in s^A for units in Ω_i^A : $t_{j,i}^L = 1$ indicates that j^{th} unit in Ω_i^A is in s_i^A and $t_{j,i}^L = 0$ indicates otherwise.

Our goal is to estimate the total Y^B , the parameter of our interest, for target population U^B which is divided into N clusters. In order to do so, we select a sample s^A from U^A with selection probability π_j^A . Then we identify Ω^B using $l_{j,ik} \neq 0$. All units of the clusters in Ω^B are surveyed where y_{ik} and the set of $l_{j,ik}$ are measured.

By applying the GWSM, an estimation weight w_{ik} will be assigned to each unit k of surveyed cluster i 's. Such weights can be chosen in an appropriate manner so that the estimator of Y^B :

$$\hat{Y}^B = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \quad (1)$$

performs well in estimating Y^B .

We are interested in estimating the quantity Y^B using \hat{Y}^B . According to Horvitz and Thompson (1952), let w_{ik} be inverse of selection probability, π_{ik} , of the k^{th} individual of U_i^B in the target population. Then \hat{Y}^B gives an unbiased estimator for Y^B . However, the computation for π_{ik} is difficult or even impossible in the present case, due to the complication in the indirect sampling scheme. Therefore, GWSM is introduced to address this issue. For readers' convenience, here we outline the GWSM in computing the weights for each cluster that has been observed.

Step 1: Provide the initial weights w'_{ik}

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}; \quad (2)$$

Step 2: Compute L_i^B

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik}; \quad (3)$$

Step 3: Obtain final weight w_i

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{L_i^B}; \quad (4)$$

Step 4: Set $w_{ik} = w_i$ for all k in i^{th} cluster.

It follows Theorem in Section 3 of Lavallée (2001) that

$$\hat{Y}^B = \sum_{i=1}^n \frac{\sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B} \sum_{k=1}^{M_i^B} y_{ik} \quad (5)$$

offers an unbiased estimator for Y^B provided all links $l_{j,ik}$ can be correctly identified. The estimation weights assigned in (5) are

$$w_{ik} = \begin{cases} \frac{\sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B}, & \text{for all units } k \text{ in cluster } i \text{ when } i \text{ in } \Omega^B; \\ 0, & \text{when } i \text{ is not in } \Omega^B. \end{cases} \quad (6)$$

A simple example is illustrated in Figure 2. We aim to estimate the total Y^B linked to the target population U^B . Suppose that we select the units $j=1$, and 2 from U^A . By selecting the unit $j=1$, we survey the units of cluster $i=1$. Likewise, by selecting the unit $j=2$, we survey the units of clusters $i=1$, and 2. We therefore have $\Omega^B = \{1, 2\}$. For each unit k of clusters i of U^B , we calculate the initial weights w'_{ik} in (2), the total number of links existing between units of U^A and units of U_i^B , L_i^B , and the final weights w_{ik} . Then, according to (5) the resulting estimator for Y^B is as below (see Lavallée 2007, pages 17-18 for more details):

$$\hat{Y}^B = \frac{1}{2} \left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right] y_{11} + \frac{1}{2} \left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right] y_{12} + \frac{1}{3\pi_2^A} y_{21} + \frac{1}{3\pi_2^A} y_{22} + \frac{1}{3\pi_2^A} y_{23}. \quad (7)$$

We note that for the estimator with known $l_{j,ik}$, the only assumption for unbiasedness is to have $L_i^B > 0$ for all clusters i 's in U^B . That is, every cluster of the target population must have at least one link from U^A . We know that if some links were missing, then the estimator (5) would be biased. When link nonresponse occurs, as indicated in Lavallée (2001), L_i^B can not be determined. Traditionally, using total links observed to replace this unknown quantity results in overestimation on Y^B since some link components are actually missing in summation L_i^B . Our proposed study focus is on just such a problem, and we attempt to adjust the estimation weights w_{ik} by estimating L_i^B so as to obtain a better performance of estimation on Y^B .

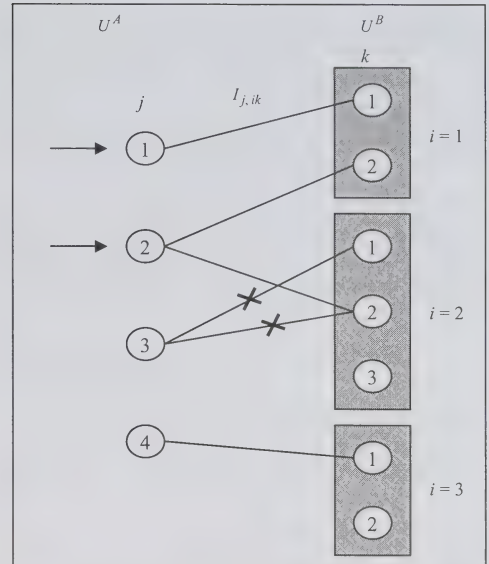


Figure 2 Example of links in indirect sampling

3. Treatments of biased estimation problems

As indicated in Section 1, the biased estimation using GWSM occurs due to link nonresponse problems. In this situation, not all of the composition in L_i^B can be identified or observed. Although the links between units in s^A and units in U^B can normally be determined in practice, the parts of links outside s^A are often difficult or even impossible to identify. We say that such units have missing links with U^B . Let $\Delta^A = \Omega^A \setminus s^A$ be the set of units with possible missing links. Then,

$$L_i^B = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta^A} \sum_{k=1}^{M_i^B} l_{j,ik}. \quad (8)$$

If we carry out the GWSM without taking these missing links into account, we use the total of observed $l_{j,ik}$ as L_i^{B*} instead to compute \hat{Y}^B using

$$L_i^{B*} = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta_0^A} \sum_{k=1}^{M_i^B} l_{j,ik}, \quad (9)$$

where Δ_0^A is a subset of Δ^A and only contains the units whose links are observed. The cost is overestimation of Y^B in using (5) since

$$L_i^B \geq L_i^{B*}.$$

We suggest a few methods for applying GWSM under consideration of link nonresponse by estimating L_i^B .

3.1 Estimating L_i^B without availability of auxiliary variables

3.1.1 Estimating L_i^B by proportional adjustment for each individual cluster (Method 1)

To address the link nonresponse problem, we focus on estimating L_i^B using the known information about the links within s^A . To compute the weights in (6) using GWSM, we only need to estimate L_i^B for those $i \in \Omega^B$. For any $i \in \Omega^B$,

$$L_i^B = \sum_{j=1}^{T_i^A} L_{j,i}^B. \quad (10)$$

A general estimator for this total can be expressed as

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} w_{j,i}^L L_{j,i}^B, \quad (11)$$

where $w_{j,i}^L$ is a random weight that takes the value $w_{j,i}^L = 0$ if j is not in the sample s^A . For each $i \in \Omega^B$, we use the known link information between s_i^A and U_i^B to estimate the link information between Ω_i^A and U_i^B . The expectation of \hat{L}_i^B is

$$E(\hat{L}_i^B) = \sum_{j=1}^{T_i^A} E(w_{j,i}^L) L_{j,i}^B. \quad (12)$$

By comparing (10) and (12), it can be observed that \hat{L}_i^B is unbiased for L_i^B for any weighting scheme with $E(w_{j,i}^L) = 1$ for all j .

First of all, we adopt the Horvitz-Thompson estimator (Horvitz & Thompson 1952), also called π estimator (Särndal, Swensson, and Wretman 1991). Note that, by the definition of Ω_i^A , $\Omega_i^A \supset s_i^A$ for all i . We imitate a procedure for estimating the number of links in Ω_i^A using that in s_i^A . The procedure is to select a "sample" s_i^A from the "population" Ω_i^A . Let $\pi_{j,i}^L$ be the probability of j (which is in Ω_i^A) being included in s_i^A . Then, let

$$w_{j,i}^L = \begin{cases} 1/\pi_{j,i}^L, & j \text{ is in } s_i^A, \\ 0, & j \text{ is in } \Omega_i^A \setminus s_i^A. \end{cases} \quad (13)$$

According to Corollary 3.1 in Cassel, Särndal, and Wretman (1977), this weighting scheme provides an unbiased estimator for L_i^B . We have

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\pi_{j,i}^L}. \quad (14)$$

It provides us with an asymptotically unbiased (proof follows) estimator of Y^B :

$$\bar{Y}^B = \sum_{i=1}^n \frac{\sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\pi_{j,i}^L}} \sum_{k=1}^{M_i^B} y_{ik}. \quad (15)$$

In order to show its unbiasedness, we employ Taylor's expansion. According to Corollary 5.1.5 (Fuller 1996), we obtain

$$\begin{aligned} \frac{1}{\hat{L}_i^B} &= \frac{1}{L_i^B} - \frac{1}{(L_i^B)^2} (\hat{L}_i^B - L_i^B) + O[(\hat{L}_i^B - L_i^B)^2] \\ &= \frac{1}{(L_i^B)^2} (2L_i^B - \hat{L}_i^B) + O_p(n^{-1}). \end{aligned}$$

It follows that

$$p \lim \left\{ n^{1/2} \left[\frac{1}{\hat{L}_i^B} - \frac{1}{(L_i^B)^2} (2L_i^B - \hat{L}_i^B) \right] \right\} = 0.$$

Therefore, by Theorem 5.2.1 (Fuller 1996), the limiting distribution of $n^{1/2}[1/\hat{L}_i^B]$ is the limiting distribution of $n^{1/2}[1/(L_i^B)^2(2L_i^B - \hat{L}_i^B)]$. We note that \bar{Y}^B is a function of both random variable: t_j , and random variable: $t_{j,i}^L$; therefore we denote the expectation of \bar{Y}^B with respect to t_j by $E_{t_j}(\cdot)$ and that with respect to $t_{j,i}^L$ by $E_{t_{j,i}^L}(\cdot)$. Hence, asymptotically we have

$$\begin{aligned} E(\bar{Y}^B) &\approx \sum_{i=1}^n E_{t_j} \left[E_{t_{j,i}^L} \left(\frac{1}{(L_i^B)^2} \left(2L_i^B - \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L} \right) \right. \right. \\ &\quad \left. \left. \sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \middle| \Omega^B \right] \sum_{k=1}^{M_i^B} y_{ik} \\ &= \sum_{i=1}^n E_{t_j} \left(\frac{1}{L_i^B} \sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \end{aligned} \quad (16)$$

$$\begin{aligned} &= E_{t_j} \left(\sum_{i=1}^n \left(\frac{1}{L_i^B} \sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \right) \\ &= E_{t_j}(\bar{Y}^B). \end{aligned} \quad (17)$$

According to Lavallée (1995), $E_{t_j}(\bar{Y}^B) = Y^B$. Therefore, \bar{Y}^B is an approximately unbiased estimator of Y^B .

Now we need to compute $\pi_{j,i}^L$. It is a function of π_j^A yet it depends on how s_i^A affects on U_i^B , therefore on Ω_i^A . Such an effect is difficult to track and varies from case to case; however, we can give a general estimate of it. The first approach we propose in this paper is to estimate selection probability, $\pi_{j,i}^L$ using the proportion of the units in s^A which take in Ω^A . Namely

$$\hat{\pi}_{j,i}^{L(1)} = \frac{m_i^A}{T_i^A}. \quad (18)$$

Therefore,

$$\begin{aligned} \hat{L}_i^{B(1)} &= \sum_{j=1}^{T^A} \frac{L_{j,i}^B t_j^L}{\hat{\pi}_{j,i}^{L(1)}} \\ &= \frac{T_i^A}{m_i^A} \sum_{j=1}^{m^A} L_{j,i}^B. \end{aligned} \quad (19)$$

and

$$\hat{Y}^{B(1)} = \sum_{i=1}^n \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\frac{T_i^A}{m_i^A} \sum_{j=1}^{m^A} L_{j,i}^B} \sum_{k=1}^{M^B} y_{ik} = \sum_{i=1}^n w_i^{(1)} \sum_{k=1}^{M^B} y_{ik}, \quad (20)$$

with

$$w_i^{(1)} = \frac{m_i^A}{T_i^A} \frac{\sum_{j=1}^{m^A} \frac{L_{j,i}^B}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B}. \quad (21)$$

We revisit the example in Figure 2, assuming that there are two link nonresponses that happened between the unit $j=3$ in U^A and the units $k=1, 2$ of cluster $i=2$ in U^B . If we use the GWSM without adjustment in (5), the resulting estimator for Y^B is no longer (7). We have instead

$$\begin{aligned} \hat{Y}^B &= \frac{1}{2} \left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{11} + \frac{1}{2} \left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{12} \\ &\quad + \frac{1}{\pi_2^A} y_{21} + \frac{1}{\pi_2^A} y_{22} + \frac{1}{\pi_2^A} y_{23}, \end{aligned} \quad (22)$$

which is biased. In order to apply (20), we first compute m_i^A/T_i^A . Then the resulting weights using Method (1) in (21) for this example is shown in Table 1. Therefore, this modified method provides the estimator:

$$\begin{aligned} \hat{Y}^B &= \frac{1}{2} \left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{11} + \frac{1}{2} \left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{12} \\ &\quad + \frac{1}{2\pi_2^A} y_{21} + \frac{1}{2\pi_2^A} y_{22} + \frac{1}{2\pi_2^A} y_{23}, \end{aligned} \quad (23)$$

which is less biased than (22).

Table 1

Initial weights, total number of responded links, and final weights from (21)

i	k	w_{ik}^*	$L_{i,k}^B$	m_i^A	T_i^A	m_i^A/T_i^A	$w_i^{(1)}$
1	1	$1/\pi_1^A$	1	2	2	1	$1/2(1/\pi_1^A + 1/\pi_2^A)$
1	2	$1/\pi_2^A$	1	2	2	1	$1/2(1/\pi_1^A + 1/\pi_2^A)$
2	1	0	0 (missing)	1	2	1/2	$1/2\pi_2^A$
2	2	$1/\pi_2^A$	1 (one link is missing)	1	2	1/2	$1/2\pi_2^A$
2	3	0	0	1	2	1/2	$1/2\pi_2^A$

3.1.2 Estimating L_i^B by overall proportional adjustment (Method 2)

In the previous approach, the information regarding m_i^A and T_i^A is needed for every i . Suppose we ignore the variation of Ω_i^A among all i , then we simply propose that

$$L_i^{B*} = \sum_{j=1}^{T^A} \frac{L_{j,i}^B t_j^L}{\pi_j^L} \quad (24)$$

using link information in s^A to estimate the link information in T^A , where t_j^L being the indicator variable for being in s^A from Ω^A . Now we need to compute π_j^L . Again it is a function of π_j^A and yet it depends on the complexity of effects of s^A on Ω^B , hence to Ω^A . While the computation is difficult and varies from case to case without a general form, we can usually give a rough estimate of it.

The second approach we propose in this paper is to estimate π_j^L using the proportion of the units in s^A which appear in Ω^A , i.e., $\pi_j^{L*} = m^A/T^A$. It informs us that

$$\hat{L}_i^{B(2)} = \frac{T^A}{m^A} \sum_{j=1}^{m^A} L_{j,i}^B. \quad (25)$$

For simple random designs with or without stratification, $\hat{L}_i^{B(2)}$ provides an unbiased estimator for L_i^B . For more complex designs, it provides a model-based unbiased estimator under assumption (A) as follows:

(A) Suppose that for any cluster i , the average of total existing links associated with all units in the sample s^A is the same as that of existing links associated with all units in U^A , i.e.,

$$\frac{\sum_{j=1}^{m^A} L_{j,i}^B}{m^A} = \frac{\sum_{j=1}^{M^B} L_{j,i}^B}{T^A}. \quad (26)$$

So, the estimation weights are provided by

$$w_{ik}^{(2)} = w_i^{(2)} = \frac{m^A \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{T^A \sum_{j=1}^{M^A} L_{j,i}^B t_j}, \text{ for all units } k \text{ in cluster } i. \quad (27)$$

It follows that Y^B can be estimated by

$$\hat{Y}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^n \sum_{j=1}^{M^A} \frac{L_{j,i}^B}{\pi_j^A} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^n w_i^{(2)} \sum_{k=1}^{M_i^B} y_{ik}, \quad (28)$$

We recall the example in Figure 2 with two link nonresponses that happened between the unit $j = 3$ in U^A and the units $k = 1, 2$ of cluster $i = 2$ in U^B . In order to apply (28), we first compute m^A/T^A . For this example, we have $m^A = 2$, and $T^A = 3$. Then the resulting estimator for Y^B using the adjustment Method (2) for this example is

$$\hat{Y}^B = \frac{2}{3} \left[\frac{1}{2} \left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{11} + \frac{1}{2} \left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{12} + \frac{1}{\pi_2^A} y_{21} + \frac{1}{\pi_2^A} y_{22} + \frac{1}{\pi_2^A} y_{23} \right]. \quad (29)$$

Therefore, this adjustment made in (28) is different from Method (1) for this example.

We know that $\text{var}(\hat{Y}^{B(1 \text{ or } 2)}) = \text{var}\{E(\hat{Y}^{B(1 \text{ or } 2)} | s^A)\} + E\{\text{var}(\hat{Y}^{B(1 \text{ or } 2)} | s^A)\}$. The inner expectation and variance (conditional on s^A) are taken over all possible sets of "responding" $l_{j,ik}$, given the sample s^A while the outer expectation and variance are taken over all possible sample s^A . Generally, the adjustments made above will not eliminate the second term which depends on the randomness of $l_{j,ik}$.

3.2 Estimating L_i^B with availability of auxiliary variables

3.2.1 Estimating $l_{j,ik}$ using logistic model

The estimation methods for L_i^B proposed in Section 3.1 are simple to apply and do not need additional information. However, sometimes the assumption can be violated which results in an undesirable estimate. For instance, $L_{j,i}^B$ may depend on some characteristics of unit j and cluster i .

We assume that the probability of a link between a unit in sampling population and a unit in target population depends on some auxiliary variables through a logistic regression model. We may estimate this probability function so that the estimation of the quantity of interest in the target population is desirable. Let $P_{j,ik} = P(l_{j,ik} = 1)$ which is

affected by some variable vector \mathbf{x}_j^A in U^A and \mathbf{x}_{ik}^B in U^B .

We may fit the logistic model

$$\log \left(\frac{P_{j,ik}}{1 - P_{j,ik}} \right) = \mathbf{a}' \mathbf{x}_j^A + \mathbf{b}' \mathbf{x}_{ik}^B \quad (30)$$

using the observed links and their corresponding characteristic variables. The unknown parameter vectors \mathbf{a} and \mathbf{b} can be estimated. Then, for those $l_{j,ik}$ s which can not be identified we suggest to impute them with their probability estimates:

$$\hat{P}_{j,ik} = \frac{e^{\hat{\mathbf{a}}' \mathbf{x}_j^A + \hat{\mathbf{b}}' \mathbf{x}_{ik}^B}}{1 + e^{\hat{\mathbf{a}}' \mathbf{x}_j^A + \hat{\mathbf{b}}' \mathbf{x}_{ik}^B}}, \quad (31)$$

where $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$ is an estimator for (\mathbf{a}, \mathbf{b}) , for instance, we use the weighted maximum likelihood (pseudolikelihood) estimator. We then have

$$\begin{aligned} \hat{L}_i^{B(3)} &= \sum_{j \in s^A \cup \Delta_0^A} L_{j,i} + \sum_{j \in \Omega^A \setminus (s^A \cup \Delta_0^A)} \hat{L}_{j,i} \\ &= \sum_{j \in s^A \cup \Delta_0^A} L_{j,i} + \sum_{j \in \Omega^A \setminus (s^A \cup \Delta_0^A)} \sum_{k=1}^{M_i^B} \frac{e^{\hat{\mathbf{a}}' \mathbf{x}_j^A + \hat{\mathbf{b}}' \mathbf{x}_{ik}^B}}{1 + e^{\hat{\mathbf{a}}' \mathbf{x}_j^A + \hat{\mathbf{b}}' \mathbf{x}_{ik}^B}}. \end{aligned} \quad (32)$$

After replacing L_i^B with $\hat{L}_i^{B(3)}$ in (5), (5) provides us with a consistent estimator for Y^B when the model specified in (30) is correct and $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$ is consistent. Note that there are alternatives for the logistic model, such as logit and complementary log-log models. See Draper and Smith (1998) for details. Their research also states that the choice of which model should be employed is not always clear in practice.

3.2.2 Directly estimating L_i^B use log-linear model

We consider that there is a variable vector \mathbf{x}_i^B which affects the value of L_i^B . This indicates that the total number of links in a cluster only varies according to the characteristics of the cluster itself. Using the log-linear model, we can propose (33) below:

$$\log(L_i^B) = \theta^T \mathbf{x}_i^B. \quad (33)$$

If the fit is reasonable, L_i^B can be estimated directly by

$$\hat{L}_i^{B(4)} = e^{\hat{\theta}^T \mathbf{x}_i^B}, \quad (34)$$

where $\hat{\theta}$ is an estimator for θ . When $\hat{\theta}$ is consistent then after replacing L_i^B with $\hat{L}_i^{B(4)}$ in (5), (5) provides a consistent estimator for Y^B . We note that $\hat{L}_i^{B(4)}$ might be non-integer valued, and therefore might have to be rounded to the nearest integer value.

4. Simulation study

When the production of cross-sectional estimates at a particular point in time after the initial point is also of interest in a longitudinal survey design, it becomes a practical example of an indirect sampling problem. Since the population changes over time, the target population is not the same as the initial population which the longitudinal sample is selected from. In this section we will use Survey of Labour and Income Dynamics (SLID) as an example to demonstrate the performance of one of the estimators we introduced in Section 3.1.

The sample design for SLID is detailed in Lavallée (1993). Some terminologies we use in this report - such as cohabitants, initially-present individuals, and initially-absent individuals - follow Lavallée (1995). Initially-absent individuals in the population are individuals who were not part of the population in the year the longitudinal sample was selected, but are considered in the later sample; included among these are newborns and immigrants. After the initial year of selection, the population contains longitudinal individuals, initially-present individuals and initially-absent individuals. Focusing on the households containing at least one longitudinal individual (*i.e.*, longitudinal households), initially-present and initially-absent individuals who join these households are referred to as cohabitants.

In this specific example, U^A is the population at the initial year, say y_0 , of the longitudinal survey, and U^B is the population at any of the following years, say year y_{r_i} , after the initial year. The sample s^A is all the longitudinal individuals. $L_{j,i}$ is a binary variable; it values 1 if individual j lives in i^{th} household at y_{r_i} ; 0 otherwise. $L_{j,i}^B$ is the total number of longitudinal persons and initially-present cohabitants at y_0 who lives in i^{th} household at y_{r_i} .

For a longitudinal individual the link would be one to one. For cohabitants there is a significant possibility that this link will be impossible to identify a few years past the initial year, for reasons such as new birth and immigration; further, the greater proportion of cohabitants occupying the target population, the larger this possibility becomes. For instance, in survey panel 3 in SLID, cohabitants represent 7.8 percents out of 47,377 individuals in the year of 2000 which is one year after the initial year. This increases to 13.87 percent in the year 2002 (3 years later), and 15.22 percent in 2003 (4 years later). We can see that the link nonresponses can not be overlooked in such a significant proportion of cohabitants. Due to the availability of observed information, we implement the approach of estimating L_i^B by two kinds of proportional adjustments, which we proposed in Section 3.1.1 and 3.1.2. In order to test the performance of the estimates obtained by these approaches, we carry out a

simulation study using SLID data. Cross-sectional estimations for four income variables are of interest for the year of 2003. These four variables are: total income before taxes; total income after taxes; earnings (includes wages and salaries before deductions and self-employment income); and wages and salaries before deductions (also called employment income). We are interested in the total of the population incomes for all these variables. These four quantities of interest have been estimated at both the national level and the provincial level.

For a longitudinal survey, the total number of links in cluster i are generally not more than the total number of individuals in this cluster and not less than the number of longitudinal individuals in this cluster. Since T_i^B is unknown, we replace T_i^B by M_i^B in (5) in our simulation study.

First, we assume that the links between all units selected in the initial year (1999) and all units in the whole population in 2003 are correctly specified. Then we compute the totals using GWSM. We use it as our estimation target, the “truth.”

Second, we randomly take away 50 percent of the links associated with initially-present individuals by setting up at random some initially present cohabitants as initially absent ones. The number of links taken makes up approximately 6.3 percent of the total population with which we are interested, with a size of 30,224. Without any adjustment, we recalculate the estimates using GWSM. We use it as our estimation benchmark, the “placebo.”

Third, we estimate the same quantities using GWSM with proportional adjustment approaches, Method (1) and (2) in Section 3.1, to see whether the estimates are close enough to the “truth” and how much improvement these adjustments make.

This simulation study using SLID data demonstrates that the proposed method performs very well in overcoming the overestimation problems that arise from link nonresponse.

We denote

$$w_i^{\text{mean}} = \frac{\sum_{j=1}^{m^A} L_{j,i}^B \frac{1}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B} \quad (35)$$

Then, using Method (1) and (2) in Section 3.1 we estimate Y^B by

$$\hat{Y}_{\text{mean}}^{B(1)} = \sum_{i=1}^n \frac{m_i^A}{T_i^A} w_i^{\text{mean}} \sum_{k=1}^{M_i^B} y_{ik}, \quad (36)$$

and

$$\hat{Y}_{\text{mean}}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^n w_i^{\text{mean}} \sum_{k=1}^{M_i^B} y_{ik}, \quad (37)$$

respectively.

We note that w_i^{mean} is the average weight of longitudinal persons who live in i^{th} household at y_{it} . Therefore, it is also reasonable to use median weight:

$$w_i^{\text{median}} = \text{the median of } \frac{1}{\pi_j^A}, j = 1, 2, \dots, m^A. \quad (38)$$

instead to enhance the robustness of the estimates. Namely, we estimate Y^B as well by

$$\hat{Y}_{\text{median}}^{B(1)} = \sum_{i=1}^n \frac{m_i^A}{T_i^A} w_i^{\text{median}} \sum_{k=1}^{M_i^B} y_{ik}, \quad (39)$$

and

$$\hat{Y}_{\text{median}}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^n w_i^{\text{median}} \sum_{k=1}^{M_i^B} y_{ik}. \quad (40)$$

The comparison for these proposed methods with and without incorporation in nonresponse problems both using mean and median weight within each household are presented in Tables 2-5.

The next four tables give the result for the performance of our estimate using relative error defined as:

$$\left| \frac{\text{estimate} - \text{"truth"}}{\text{"truth"}} \right| \times 100\%.$$

Table 2
Total income before taxes (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	9,261,958,108	9,788,749,735	9,317,420,236	9,304,530,248
PEI	2,720,448,008	2,858,506,466	2,735,943,043	2,734,922,451
NS	18,277,017,251	19,573,546,299	18,140,076,618	18,067,144,557
NB	15,297,155,323	16,281,178,934	15,291,696,585	15,236,482,035
QC	1.57839E+11	1.69664E+11	1.56533E+11	1.56405E+11
ON	2.895E+11	3.07642E+11	2.85409E+11	2.85599E+11
MA	23,436,397,548	25,043,168,032	23,632,717,226	23,553,543,216
SK	20,185,285,649	21,595,804,296	20,163,683,598	20,095,359,071
AB	69,063,402,292	74,576,351,600	68,716,661,193	68,582,541,733
BC	81,749,374,346	86,593,614,506	81,387,640,982	81,248,680,715
National	6.8733E+11	7.33617E+11	6.8286E+11	6.82356E+11

Table 3
Total income after taxes (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	7,846,587,557	8,287,351,908	7,892,754,014	7,882,437,105
PEI	2,300,092,795	2,416,503,441	2,314,256,124	2,313,544,320
NS	15,154,508,564	16,257,679,161	15,080,155,194	15,020,088,623
NB	12,878,350,198	13,718,260,686	12,894,700,593	12,849,252,205
QC	1.27632E+11	1.37514E+11	1.27118E+11	1.26999E+11
ON	2.3788E+11	2.53073E+11	2.35192E+11	2.3534E+11
MA	19,541,510,220	20,877,377,918	19,713,628,649	19,649,142,217
SK	16,894,929,025	18,073,635,883	16,890,410,993	16,834,787,407
AB	57,466,974,767	62,055,315,246	57,183,814,491	57,073,904,623
BC	68,710,569,670	72,770,595,462	68,431,531,373	68,309,055,749
National	5.66306E+11	6.05044E+11	5.63958E+11	5.63518E+11

Table 4
Earnings (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	6,433,112,169	6,837,522,157	6,541,306,193	6,530,174,122
PEI	1,898,192,704	2,019,341,995	1,964,066,449	1,962,669,664
NS	12,772,667,160	13,809,197,160	12,999,111,234	12,939,785,579
NB	11,250,688,811	12,030,378,710	11,411,530,716	11,370,222,533
QC	1.18878E+11	1.28949E+11	1.19797E+11	1.19717E+11
ON	2.27577E+11	2.43404E+11	2.26812E+11	2.27092E+11
MA	17,560,695,670	18,995,682,322	18,066,353,153	18,001,882,362
SK	15,159,319,031	16,340,668,148	15,381,733,004	15,319,210,228
AB	56,152,023,359	61,059,244,608	56,540,145,524	56,418,889,147
BC	60,532,655,979	64,499,398,960	61,192,920,832	61,085,986,951
National	5.28214E+11	5.67945E+11	5.3199E+11	5.31722E+11

Table 5
Wages and salaries before deductions (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	6,180,713,343	6,572,345,010	6,283,079,555	6,272,429,515
PEI	1,636,344,440	1,747,755,878	1,713,809,312	1,713,157,676
NS	12,327,220,137	13,341,912,666	12,579,519,733	12,521,159,025
NB	10,742,381,379	11,508,445,078	10,961,105,589	10,921,102,477
QC	1.08636E+11	1.18092E+11	1.10024E+11	1.09898E+11
ON	2.07331E+11	2.22043E+11	2.07265E+11	2.07495E+11
MA	16,146,993,217	17,504,024,442	16,701,823,718	16,641,840,086
SK	13,982,423,360	15,129,217,320	14,311,467,435	14,255,519,224
AB	52,594,490,290	57,359,188,114	53,195,227,508	53,077,388,907
BC	56,206,787,033	59,886,429,369	56,875,663,895	56,764,297,512
National	4.85784E+11	5.23184E+11	4.91116E+11	4.90763E+11

Table 6
Comparison of relative errors in estimating income before taxes (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	5.688	0.599	0.460	1.059	2.397
PEI	5.075	0.570	0.532	2.859	4.063
NS	7.094	0.749	1.148	3.549	2.459
NB	6.433	0.037	0.397	2.693	2.987
QC	7.492	0.828	0.909	4.372	2.896
ON	6.267	1.413	1.348	4.691	1.771
MA	6.856	0.838	0.500	1.644	3.654
SK	6.988	0.107	0.446	2.480	2.598
AB	7.982	0.502	0.696	3.185	2.407
BC	5.926	0.442	0.612	3.995	3.343
National	6.734	0.650	0.724	3.868	2.662

Table 7
Comparison of relative errors in estimating income after taxes (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	5.617	0.588	0.457	1.101	2.409
PEI	5.061	0.616	0.585	2.832	4.121
NS	7.279	0.491	0.887	3.338	2.765
NB	6.522	0.127	0.226	2.539	3.150
QC	7.742	0.403	0.496	3.991	3.375
ON	6.387	1.130	1.068	4.432	2.081
MA	6.836	0.881	0.551	1.645	3.733
SK	6.977	0.027	0.356	2.406	2.675
AB	7.984	0.493	0.684	3.180	2.415
BC	5.909	0.406	0.584	3.989	3.419
National	6.841	0.415	0.492	3.657	2.927

Table 8
Comparison of relative errors in estimating earnings (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	6.286	1.682	1.509	0.041	3.585
PEI	6.382	3.470	3.397	0.0739	7.115
NS	8.115	1.773	1.308	1.265	5.281
NB	6.930	1.430	1.062	1.279	4.512
QC	8.472	0.773	0.706	2.827	4.560
ON	6.955	0.336	0.213	3.760	2.920
MA	8.172	2.879	2.512	0.291	5.835
SK	7.793	1.467	1.055	0.979	4.324
AB	8.739	0.691	0.475	2.140	3.777
BC	6.553	1.091	0.914	2.643	5.081
National	7.522	0.715	0.664	2.628	4.131

They show that our estimates using both method (1) and method (2) perform very well in terms of reducing bias. Method (1) does work better than Method (2) overall, yet the improvement from Method (1) to Method (2) is much less compared to that made by moving from without adjustment to method (2). Since Method (2) provides us with high quality and involves much less information than Method (1), Method (2) is recommended.

Now, we focus on Method (2) using mean, which gives the estimate $\hat{Y}_{\text{mean}}^{B(2)}$ to analyze how its variance performs in terms of estimating Y^B . We use the bootstrap technique to estimate the variance of $\hat{Y}_{\text{mean}}^{B(2)}$ at both the national level and the provincial level. The bootstrap used for our simulation in this paper is the classical Bootstrap with replacement, where bootstrapping is performed at the first stage of sampling. The bootstrap weights taken here are provided with the SLID data, and incorporate all the necessary adjustments. See Lévesque (2001), and LaRoche (2003) for details on the use of the Bootstrap for SLID. The improvement in

reducing the variance is not as large as in reducing bias; however, it is revealed in this simulation study that the proposed method provides a smaller variance as well compared to applying GWSM without an adjustment for missing links. See Table 10 for the results.

The simulation results presented here are based on a single sample of SLID and a single random removal of the links of initially-present individuals. For a complete assessment of the properties of the above estimators, a Monte-Carlo process would have been suitable. Such simulations have been performed by Hurand (2006) based on agricultural data. In these simulations, 1,000 samples have been selected and for each selected sample, the worst-case-scenario has been used, *i.e.*, all links from the non-sample units have been removed. The results of these simulations showed that proportional adjustment and global proportional adjustment are the two methods whose estimates are, on average, the closest to the real total, and whose biases are negligible.

Table 9
Comparison of relative errors in estimating wages and salaries before deductions (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	6.336	1.656	1.484	0.1012	3.593
PEI	6.809	4.734	4.694	1.056	8.424
NS	8.231	2.047	1.573	0.939	5.509
NB	7.131	2.036	1.664	0.685	5.133
QC	8.704	1.278	1.162	2.294	5.070
ON	7.096	0.0317	0.0791	3.473	3.265
MA	8.404	3.436	3.065	0.787	6.469
SK	8.202	2.353	1.953	0.107	5.213
AB	9.059	1.142	0.918	1.713	4.247
BC	6.547	1.190	0.992	2.565	5.234
National	7.699	1.098	1.025	2.251	4.541

Table 10
Comparison of standard deviation estimates

	Variables	Total income before taxes	Total income after taxes	Earnings	Wages and salaries before deductions
National	GWSM with missing links	9,677,258,789	7,343,792,762	8,850,202,075	8,468,718,449
level	Method (2) using mean	9,471,103,083	7,238,715,323	8,593,015,854	8,232,428,642
Ontario	GWSM with missing links	7,888,106,377	6,101,001,739	7,245,688,373	7,149,203,530
	Method (2) using mean	7,601,169,501	5,939,509,894	6,952,217,872	6,831,300,511
Quebec	GWSM with missing links	4,341,215,711	3,113,247,130	3,772,369,180	3,162,277,660
	Method (2) using mean	4,160,251,472	2,974,248,451	3,668,996,929	3,100,868,366

5. Closing remarks

We have constructed four estimation methods to address the link nonresponse problem in indirect sampling. The simulation results in this article show that the adjustments methods we have presented in the example for using GWSM incorporating the link nonresponse performs well in terms of both reducing the estimation bias and providing an overall improvement in variance. The advancement in bias reduction seems significant. The implementation of the methods proposed in Section 3.2 for real data sets will be studied in the near future.

The following significant observations emerged from our study:

1. Adjustment methods are simple to apply.
2. In a more general situation, such as $L_{j,i} > 1$ for some j 's, (35) represents the weighted mean weighted by $L_{j,i}^B$. Accordingly the median approach delivered by (39) and (40) can be modified using a generalized version of median – “weighted” median. Namely, we replace (38) by

$$w_i^{\text{median}} = \text{the median of } \frac{1}{\pi_j^A}$$

where $j = 1, 2, \dots, L_{1,i}^B; 1, 2, \dots, L_{2,i}^B; \dots; 1, 2, \dots, L_{m^A,i}^B$.

3. Some valid link responses outside s^A can not be used in estimating L_i^B by the methods proposed in Section 3.1. However, this valid information would be beneficial to the approaches by predicting $l_{j,ik}$ using auxiliary variables, as can be seen in Section 3.2.1.

Acknowledgements

The authors would like to thank the Associate Editor and the two referees for their helpful suggestions and comments on the previous versions of this paper. This research is funded by the Natural Sciences and Engineering Research Council of Canada, and Mathematics of Information Technology and Complex Systems.

References

- Cassel, C.-M., Särndal, C.-E. and Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons, Inc.

- Déville, J.-C., and Lavallée, P. (2006). Indirect sampling: Foundations of the generalised weight share method. *Survey Methodology*, 32, 165-176.
- Draper, N.R., and Smith, H. (1998). *Applied Regression Analysis*, 3rd Ed. New York: John Wiley & Sons, Inc.
- Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc., 135-159.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Hurand, C. (2006). La méthode généralisée du partage des poids et le problème d'identification des liens. Internal report of the Social Survey Methods Division, Statistics Canada, July 2006.
- Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- LaRoche, S. (2003). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics. *Income Research Paper Series*, Catalogue no. 75F0002MIE - No. 007, Statistics Canada.
- Lavallée, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. *Statistics Canada, Research Paper of the Survey of Labour and Income Dynamics*, Catalogue No. 93-19, December 1993.
- Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using weight share method. *Survey Methodology*, 21, 25-32.
- Lavallée, P. (2001). Correcting for non-response in indirect sampling. *Proceedings of Statistics Canada's Symposium 2001*.
- Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles and Éditions Ellipse.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.
- Lévesque, I. (2001). Enquête sur la dynamique du travail et du revenu - Estimation de la variance. Internal document from Statistics Canada, July 2, 2001.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Nonparametric propensity weighting for survey nonresponse through local polynomial regression

Damião N. da Silva and Jean D. Opsomer¹

Abstract

Propensity weighting is a procedure to adjust for unit nonresponse in surveys. A form of implementing this procedure consists of dividing the sampling weights by estimates of the probabilities that the sampled units respond to the survey. Typically, these estimates are obtained by fitting parametric models, such as logistic regression. The resulting adjusted estimators may become biased when the specified parametric models are incorrect. To avoid misspecifying such a model, we consider nonparametric estimation of the response probabilities by local polynomial regression. We study the asymptotic properties of the resulting estimator under quasi-randomization. The practical behavior of the proposed nonresponse adjustment approach is evaluated on NHANES data.

Key Words: Kernel regression; Missing data; Propensity scores; Unit nonresponse; Weighting adjustment.

1. Introduction

Propensity weighting is a procedure that is often applied in sampling surveys to compensate for unit nonresponse. Under this type of nonresponse, complete data collection is accomplished at only a part of the units selected to the sample, which are termed as the respondents. The propensity weighting procedure operates by increasing the sampling weights of the respondents in the sample using estimates of the probabilities that they responded to the survey. These probabilities are also referred to as response propensities in virtue of their analogy with the propensity score theory of Rosenbaum and Rubin (1983) for observational studies, incorporated into survey nonresponse problems by David, Little, Samuël and Triest (1983).

General descriptions of propensity weighting to adjust classical survey estimators for nonresponse can be seen, for example, in Nargundkar and Joshi (1975), Cassel, Särndal and Wretman (1983) and Groves, Dillman, Eltinge and Little (2002). Traditionally, the way the procedure is implemented estimates the response probabilities with parametric regression curves, such as logistic, probit or exponential models. See Alho (1990), Folsom (1991), Ekholm and Laaksonen (1991) and Iannacchione, Milne and Folsom (1991) for earlier references. A recent theoretical account of the statistical properties of the procedure is given in Kim and Kim (2007). These parametric models are readily fitted as generalized linear models. However, an important and sometimes overlooked part of this procedure is the specification of the form of the link function to relate the response propensities and a linear predictor of the auxiliary information. If this function, which we shall refer to as the response propensity function, is misspecified, the resulting adjusted estimators of the population quantities are likely to be biased.

Another approach to estimate the response propensities is through nonparametric methods. The main motivation to use such methods is that the parametric form for the response propensity function need not be specified. In this sense, these methods offer an appealing alternative to the choice of a link function, as raised by Laaksonen (2006), or when a parametric model is difficult to specify a priori. In this context, Giommi (1984) proposed using kernel smoothing, in the form of the Nadaraya-Watson estimator, to estimate the response probabilities. Da Silva and Opsomer (2006) established the consistency of Giommi's estimator for the population mean and derived rates for the asymptotic bias and the variance. Theoretical properties of a Jackknife variance estimator were also studied.

In this article, we extend the results of Da Silva and Opsomer (2006) in two directions. First, we consider the estimation of the response propensities by local polynomial regression, a nonparametric technique described, for instance, in Wand and Jones (1995). Compared to kernel smoothing, local polynomial regression improves the local approximation to the unknown propensity function, which results in better practical and theoretical properties. It is also much more prevalent as a smoothing method in practice, with implementations available in most major statistical programs. Second, we apply the nonparametric propensity score estimation approach to data from the National Health and Nutrition Examination Survey (NHANES), which makes it possible to compare several nonresponse adjustment methods, both parametric and nonparametric, in a realistic setting.

In Section 2, we introduce the weighting procedure and the estimation of the response propensities. The theoretical properties of the adjusted estimators are discussed in Section 3. In section 4, we describe how to adapt a replication

1. Damião N. da Silva, Departamento de Estatística, Campus Universitário, Natal, RN 59078-970, Brazil. E-mail: damiao@ccet.ufm.br; Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, U.S.A. E-mail: jopsomer@stat.colostate.edu.

variance procedure to estimate the variance of the proposed adjusted estimators. Finally, in Section 5, we demonstrate the finite sample properties of the estimators by means of a simulation experiment using data from NHANES.

2. Weighting by local polynomial regression

Consider a population of N_v units, denoted by $U_v = \{1, 2, \dots, N_v\}$. Suppose that a sample s_v is drawn from U_v , according to some probabilistic sampling design $p(s_v)$. Let n_v be the size of s_v and $\pi_i = \pi_{i_v} = \Pr\{i \in s_v\} = \sum_{s_v: i \in s_v} p(s_v)$ be the inclusion probability of unit i , for all $i \in U_v$. It is of interest to estimate the population mean of a study variable y , namely $\bar{y}_{N_v} = N_v^{-1} \sum_{i \in U_v} y_i$, where y_i denotes the value of y for the i^{th} unit of U_v . We assume that the values x_i of an auxiliary variable x are fully observed throughout the sample. Let $y_v = (y_1, \dots, y_{N_v})$, and similarly for x_v .

When the sample contains unit nonresponse, we only observe the values of the study variables for the units in a subset $r_v \subset s_v$. To account for the information lost in the estimation of the parameters of interest, it becomes necessary to model the response process. To define this response model, let R_i be an indicator variable assuming the value one if the unit i respond to the survey, and the value zero otherwise, for all $i \in s_v$. We assume that, given the sample, the response indicators are independent Bernoulli random variables with

$$\Pr\{R_i = 1 \mid i \in s_v, y_v, x_v\} = \phi(x_i) \equiv \phi_i, \text{ for all } i \in s_v, \quad (1)$$

where the exact form of the *response propensity function* $\phi(\cdot)$ is unspecified, but it is assumed to be a smooth function of x_i with $\phi(\cdot) \in (0, 1]$. The relationship in (1) defines a nonresponse process said to be *ignorable*, in the sense that the response propensities are independent of the values of any study variable, conditional on the covariate x (see Lohr 1999, page 265). The theory developed here, therefore, does not intend to handle non-ignorable response mechanisms.

If all response propensities were known, resulting weighting adjustments could be obtained by applying a two-phase estimation approach. For instance, two possible estimators of the population mean \bar{y}_{N_v} would be given by

$$\bar{y}_{\pi\phi v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} y_i R_i \quad (2)$$

and

$$\bar{y}_{\text{rat}, \pi\phi v} = \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} y_i R_i / \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} R_i, \quad (3)$$

which are forms of adjustments for the Horvitz-Thompson and the Hájek estimators to compensate for the unit nonresponse. The same ideas can be used to obtain propensity weighting adjustments for the generalized regression estimator for estimation in the presence of nonresponse (Cassel *et al.* 1983).

Estimators (2) and (3) are unbiased and nearly unbiased for \bar{y}_{N_v} respectively, under the quasi-randomization approach of Oh and Scheuren (1983), where the statistical properties are evaluated using the joint distribution of the sampling design and the response model. However, the response propensities are usually unknown in practice and we need to replace the ϕ_i in (2) and (3) by estimates $\hat{\phi}_i$, satisfying $0 < \hat{\phi}_i \leq 1$. The resulting propensity weighting estimators are therefore

$$\bar{y}_{\pi\hat{\phi}v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} y_i R_i \quad (4)$$

and

$$\bar{y}_{\text{rat}, \pi\hat{\phi}v} = \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} y_i R_i / \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} R_i. \quad (5)$$

The latter formula has the advantage of being location-scale invariant, because the summation of its adjusted weights $\pi_i^{-1} \hat{\phi}_i^{-1} R_i / \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} R_i$ is equal to one, and does not require the population size N_v to be known.

In order to implement the propensity weighting estimators (4) and (5), it is necessary to estimate the response propensities $\hat{\phi}_i$. Da Silva and Opsomer (2006) used kernel regression for this purpose. The procedure we consider here is local polynomial regression, which can be described as follows. Let $K(\cdot)$ be a continuous and positive kernel function and h_v be its bandwidth. Define the $N_v \times (k+1)$ matrix

$$\mathbf{X}_{U_i} = \begin{bmatrix} 1 & (x_1 - x_i) & \cdots & (x_1 - x_i)^k \\ \vdots & \vdots & & \vdots \\ 1 & (x_{N_v} - x_i) & \cdots & (x_{N_v} - x_i)^k \end{bmatrix},$$

the $N_v \times N_v$ matrix

$$\mathbf{W}_{U_i} = \text{diag} \left\{ \frac{1}{h_v} K \left(\frac{x_j - x_i}{h_v} \right) : 1 \leq j \leq N_v \right\}.$$

and population vector of response indicators $\mathbf{R}_{U_i} = (R_1, R_2, \dots, R_{N_v})'$. The vector \mathbf{R}_{U_i} would be known if, instead of the sample s_v , a census was considered from the population U_v . In that case, the local polynomial regression estimator of degree k of $\phi_i = \phi(x_i)$, based on the whole population, would be given by the fit

$$\hat{\phi}_{U_i} = \mathbf{e}'_i (\mathbf{X}'_{U_i} \mathbf{W}_{U_i} \mathbf{X}_{U_i})^{-1} \mathbf{X}'_{U_i} \mathbf{W}_{U_i} \mathbf{R}_{U_i}, \quad (6)$$

where \mathbf{e}_j denotes the j^{th} column of the identity matrix of order $k+1$ and it is assumed that $\mathbf{X}'_{U_i} \mathbf{W}_{U_i} \mathbf{X}_{U_i}$ is non-singular.

Since the values of the response indicators are only observed for those units selected into the sample, the population fit (6) is unfeasible. However, defining \mathbf{X}_{s_i} as the $n_v \times (k+1)$ matrix formed with the rows of \mathbf{X}_{U_i} corresponding to the units $j \in s_v$,

$$\mathbf{W}_{s_i} = \text{diag} \left\{ \frac{1}{\pi_j h_v} K \left(\frac{x_j - x_i}{h_v} \right) : j \in s_v \right\}$$

and $\mathbf{R}_{s_i} = (\mathbf{R}_j : j \in s_v)'$, then a sample-based local polynomial regression estimator of degree k of $\phi_i = \phi(x_i)$ is given by

$$\hat{\phi}_i^o = \mathbf{e}'_i \hat{\mathbf{T}}_{s_i}^{-1} \hat{\mathbf{t}}_{s_i} \quad (7)$$

where

$$\begin{aligned} (\hat{\mathbf{T}}_{s_i}, \hat{\mathbf{t}}_{s_i}) &\equiv (\{\hat{T}_{si, pq}\}_{p,q=1}^{k+1}, \{\hat{t}_{si, p}\}_{p=1}^{k+1}) \\ &= (\mathbf{X}'_{s_i} \mathbf{W}_{s_i} \mathbf{X}_{s_i}, \mathbf{X}'_{s_i} \mathbf{W}_{s_i} \mathbf{R}_{s_i}) \end{aligned}$$

and it is assumed that $\hat{\mathbf{T}}_{s_i}$ is invertible. An special case of (7) is obtained by considering $k = 0$, which corresponds to the kernel regression estimator of Da Silva and Opsomer (2006). Other special cases from (7) are the local linear, the local quadratic and the local cubic response propensity estimators, which result from the local fit of polynomials of degree one, two and three, respectively.

In practice, when $\hat{\mathbf{T}}_{s_i}$ happens to be singular, a simple procedure to insure that $\hat{\phi}_i^o$ is well defined is choosing a bandwidth large enough to guarantee at least $k+1$ values of R_j in the window $[x_i - h_v, x_i + h_v]$, for all $i \in s_v$. If this window does not contain enough responses indicators and the bandwidth has to remain fixed, another approach has to be considered. To this purpose, we adopt here the adjustment made by Breidt and Opsomer (2000) and define the sample-based local polynomial regression estimator of degree k of $\phi_i = \phi(x_i)$ by

$$\hat{\phi}(x_i, k, h_v) = \mathbf{e}'_i \left(\hat{\mathbf{T}}_{s_i} + \text{diag} \left\{ \frac{\delta_1}{N_v} \right\} \right)^{-1} \hat{\mathbf{t}}_{s_i}, \quad i \in s_v. \quad (8)$$

where δ_1 is some small positive constant. The smaller order terms δ_1/N_v added to the main diagonal of $\hat{\mathbf{T}}_{s_i}$ are sufficient to make the resulting adjusted matrix invertible for any h_v . As a consequence, $\hat{\phi}(x_i, k, h_v)$ will be well defined, for all $i \in s_v$. However, another technical difficulty to use $\hat{\phi}(x_i, k, h_v)$ as a propensity weighting adjustment arises because the response propensity estimator (8) can indeed become arbitrarily close to zero. To tackle

this problem, we bound $\hat{\phi}(x_i, k, h_v)$ away from zero by considering the estimator

$$\hat{\phi}_i = \max \{ \hat{\phi}(x_i, k, h_v), \delta_2 (N_v h_v)^{-1} \}, \quad (9)$$

for some constant $\delta_2 > 0$. This idea is related to the adjustment made by Da Silva and Opsomer (2006) for the kernel regression estimator.

3. Asymptotic properties

In this section, we present the properties of the propensity weighting estimators (4) and (5) under estimation of the response propensities by the local polynomial estimator (9). The assumptions, lemmas and outlines of the proofs for the following results are given in the Appendix, and a complete theoretical investigation can be found in Da Silva and Opsomer (2008). The full derivations are not reported in this article, because they follow the general approach described in Da Silva and Opsomer (2006). We consider an asymptotic framework by which the population U_v is embedded into the increasing sequence of populations $\{U_v : N_v < N_{v+1}\}_{v=1}^{\infty}$. From each U_v , a sample s_v of size $n_v (n_v \geq n_{v-1})$ is selected according to a sampling design $p_v(\cdot)$. This framework is commonly adopted in asymptotic studies of survey estimators. See Isaki and Fuller (1982) for an early reference.

As a population-based approximation for $\phi_i \equiv \phi(x_i)$, we shall consider in the derivation of most results in this section the population fit by local polynomial regression

$$\tilde{\phi}_i \equiv \tilde{\phi}(x_i, k, h_v) = \mathbf{e}'_i \mathbf{B}_i \equiv \mathbf{e}'_i \mathbf{T}_i^{-1} \mathbf{t}_i, \quad i \in U_v, \quad (10)$$

where

$$\begin{aligned} (\mathbf{T}_i, \mathbf{t}_i) &\equiv (\{T_{i, pq}\}_{p,q=1}^{k+1}, \{t_{i, p}\}_{p=1}^{k+1}) \\ &\equiv E(\hat{\mathbf{T}}_{s_i}, \hat{\mathbf{t}}_{s_i}) = (\mathbf{X}'_{U_i} \mathbf{W}_{U_i} \mathbf{X}_{U_i}, \mathbf{X}'_{U_i} \mathbf{W}_{U_i} \phi_U), \end{aligned}$$

the matrices \mathbf{X}_{U_i} and \mathbf{W}_{U_i} are as in (6) and $\phi_U = (\phi(x_1), \phi(x_2), \dots, \phi(x_{N_v}))'$. The following theorem states the asymptotic properties of $\bar{y}_{\pi\hat{\phi}_v}$ under a set of assumptions in the Appendix. These assumptions are regularity conditions on the sampling design and the finite population, both of which are standard infinite population asymptotics, ignorability conditions on the nonresponse mechanism, and a set of standard regularity conditions related to the local polynomial regression of the response propensity function.

Theorem 1. Assume the assumptions (A1)-(A4), (B1)-(B3) and (C1)-(C5) in the Appendix hold. Consider the estimation of the population mean \bar{y}_{N_v} by the propensity weighting estimator $\bar{y}_{\pi\hat{\phi}_v}$ defined in (4), and suppose the response propensities are estimated by $\hat{\phi}_i$, the local polynomial regression estimator of degree k in (9). Let

$$\bar{y}_{\pi\hat{\psi}_v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \hat{\psi}_i^{-1} y_i R_i, \quad (11)$$

where

$$\hat{\psi}_i^{-1} = \tilde{\phi}_i^{-1} - \tilde{\phi}_i^{-2} \mathbf{e}_i' \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si} \mathbf{B}_i),$$

$\hat{\mathbf{t}}_{si}$ and $\hat{\mathbf{T}}_{si}$ are given in (7) and $\tilde{\phi}_i$, \mathbf{B}_i , \mathbf{T}_i are defined in (10). Then,

$$E[(\bar{y}_{\pi\hat{\psi}_v} - \bar{y}_{\pi\psi_v})^2] = O\left(\frac{1}{n_v^2 h_v^2}\right) \quad (12)$$

and the bias and variance of $\bar{y}_{\pi\hat{\psi}_v}$ satisfy

$$B_v \equiv E[\bar{y}_{\pi\hat{\psi}_v} - \bar{y}_{N_v}] = \begin{cases} O(h_v^{k+(3/2)}) + O\left(\frac{1}{n_v h_v}\right) & k \text{ even,} \\ O(h_v^{k+1}) + O\left(\frac{1}{n_v h_v}\right) & k \text{ odd,} \end{cases} \quad (13)$$

and

$$\text{Var}[\bar{y}_{\pi\hat{\psi}_v}] = O\left(\frac{1}{n_v h_v}\right). \quad (14)$$

Results (12) and (13) imply that the propensity weighting estimator $\bar{y}_{\pi\hat{\psi}_v}$, using a response propensity estimator based on local polynomial regression, is asymptotically unbiased for the population mean \bar{y}_{N_v} under the joint distribution of the sampling design and the response model (1). Combining this result with (14), then we obtain that

$$\hat{y}_{\pi\hat{\psi}_v} = \bar{y}_{N_v} + O_p\left(\frac{1}{\sqrt{n_v h_v}}\right), \quad (15)$$

when the bandwidth satisfies

$$h_v = \begin{cases} O\left(n_v^{-\frac{1}{2k+4}}\right), & k \text{ even,} \\ O\left(n_v^{-\frac{1}{2k+3}}\right), & k \text{ odd.} \end{cases} \quad (16)$$

Hence, without assuming a parametric form for the response propensity function $\phi(\cdot)$, $\bar{y}_{\pi\hat{\psi}_v}$ is consistent for the population mean with respect to the sampling design and the response model, as long as the response propensities are a smooth function of the covariate x . As a price paid for this robustness, the rate of convergence is of order $\sqrt{n_v h_v}$ instead of the usual parametric rate $\sqrt{n_v}$. However, as the degree of the local polynomial k increases, the rate of convergence improves. Since the kernel regression estimator in Da Silva and Opsomer (2006) is equivalent to the

case $k = 0$, local polynomial regression with higher degree is asymptotically superior to kernel regression in the context of a nonresponse adjustment. This theoretical finding is consistent with that in other contexts (see *e.g.*, Wand and Jones 1995, page 130).

Expression (11) on Theorem 1 generalizes another finding from Da Silva and Opsomer (2006) to the case of local polynomial regression, which is that the asymptotic weights $\hat{\psi}_i^{-1}$ cannot be approximated by the inverse of response propensities ϕ_i^{-1} (or their population-level estimators $\tilde{\phi}_i^{-1}$). One immediate consequence is that the estimator $\bar{y}_{\pi\hat{\psi}_v}$ is *not* asymptotically equivalent to $\bar{y}_{\pi\phi_v}$ in (2).

The following corollary provides an asymptotic distribution for $\bar{y}_{\pi\hat{\psi}_v}$, assuming the asymptotic normality of $\bar{y}_{\pi\psi_v}$.

Corollary 1. Assume the conditions of Theorem 1 hold. Suppose that the sampling design and the response model are such that

$$\frac{\bar{y}_{\pi\hat{\psi}_v} - \bar{y}_{N_v} - B_v}{[\text{Var}(\bar{y}_{\pi\hat{\psi}_v})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } v \rightarrow \infty,$$

where B_v is defined in (13). If additionally

$$\lim_{v \rightarrow \infty} (n_v h_v) \text{Var}(\bar{y}_{\pi\hat{\psi}_v}) \in (0, \infty),$$

then

$$\frac{\bar{y}_{\pi\hat{\psi}_v} - \bar{y}_{N_v} - B_v}{[\text{Var}(\bar{y}_{\pi\hat{\psi}_v})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

We now discuss the properties of the ratio-based version of propensity weighting estimator given in (5). Based on the results for $\bar{y}_{\pi\hat{\psi}_v}$, standard ratio estimation theory can be used to derive asymptotic results for $\bar{y}_{\text{rat}, \pi\hat{\psi}_v}$. In particular, under the same assumptions the asymptotic rates for the approximate bias and variance of $\bar{y}_{\text{rat}, \pi\hat{\psi}_v}$ are the same as those in Theorem 1, and the asymptotic distribution of $\bar{y}_{\text{rat}, \pi\hat{\psi}_v}$ is given in the following result.

Theorem 2. Assume the conditions of Theorem 1 hold. Suppose the population mean is to be estimated by the propensity weighted estimator $\bar{y}_{\text{rat}, \pi\hat{\psi}_v}$ of (5) and the response propensities are estimated by $\hat{\phi}_i$, the local polynomial regression estimator of degree k defined in (8). Let

$$\bar{e}_{\pi\hat{\psi}_v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \hat{\psi}_i^{-1} (y_i - \bar{y}_{N_v}) R_i,$$

where the weights $\hat{\psi}_i^{-1}$ are given in Theorem 1. Suppose that

$$\frac{\bar{e}_{\pi\hat{\psi}_v} - E(\bar{e}_{\pi\hat{\psi}_v})}{[\text{Var}(\bar{e}_{\pi\hat{\psi}_v})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } v \rightarrow \infty,$$

and

$$\lim_{v \rightarrow \infty} (n_v, h_v) \text{Var}(\bar{e}_{\pi \hat{\psi}_v}) \in (0, \infty).$$

Then,

$$\frac{\bar{y}_{\text{rat}, \pi \hat{\psi}_v} - \bar{y}_{N_v} - B_{\text{rat}, v}}{[\text{Var}(\bar{e}_{\pi \hat{\psi}_v})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

as $v \rightarrow \infty$, where $B_{\text{rat}, v} = O(h_v^{k+1})$, if k is odd, and $B_{\text{rat}, v} = O(h_v^{k+(3/2)})$, if k is even.

4. Variance estimation

As noted in Section 3, the estimator $\bar{y}_{\pi \hat{\psi}_v}$ is not asymptotically equivalent to $\bar{y}_{\pi \hat{\psi}_v}$, so that approximating the asymptotic variance of the former by that of the latter is typically incorrect. In fact, a proof that the asymptotic variance of $\bar{y}_{\pi \hat{\psi}_v}$ overestimates the variance of $\bar{y}_{\pi \hat{\psi}_v}$ is given by Kim and Kim (2007) when the response propensities are assumed to follow a parametric model. In the present context, the asymptotic variance of $\bar{y}_{\pi \hat{\psi}_v}$ is

$$\text{Var}[\bar{y}_{\pi \hat{\psi}_v}] = \text{Var}\left(\frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \hat{\psi}_i^{-1} R_i y_i\right),$$

with $\hat{\psi}_i^{-1}$ given in Theorem 1. As was previously noted in Da Silva and Opsomer (2006) for the simpler case of a zero degree polynomial, the high level of complexity in the expression makes direct estimation of this variance impractical, and a replication method was proposed instead. We briefly outline the procedure here, which is extended to local polynomials of degree k . We omit the theoretical derivations.

We start from a set of replicate weights in the absence of nonresponse, defined for estimating the variance of a linear estimator

$$\hat{\theta} = \frac{1}{N_v} \sum_{i \in s_v} w_i y_i.$$

The replicate variance estimator for $\hat{\theta}$ is defined as

$$\hat{V}(\hat{\theta}) = \sum_{\ell=1}^{L_v} c_\ell (\hat{\theta}^{(\ell)} - \hat{\theta})^2, \quad (17)$$

where

$$\hat{\theta}^{(\ell)} = \frac{1}{N_v} \sum_{i \in s_v} w_i^{(\ell)} y_i, \quad \ell = 1, 2, \dots, L_v,$$

denotes a set of L_v replicates for $\hat{\theta}$, $w_i^{(\ell)}$ are sampling weights associated with the ℓ^{th} replicate and c_ℓ is factor that depends on the replication procedure. Examples of replication procedures satisfying (17) use variants of the

Jackknife method or the Balanced Repeated Replication technique. The process to adapt the replication procedure to estimating the variance of $\bar{y}_{\pi \hat{\psi}_v}$ and $\bar{y}_{\text{rat}, \pi \hat{\psi}_v}$ is straightforward. The needed replicates of these adjusted estimators, namely $\bar{y}_{\pi \hat{\psi}_v}^{(\ell)}$ and $\bar{y}_{\text{rat}, \pi \hat{\psi}_v}^{(\ell)}$, are obtained by replacing the $w_i = \pi_i^{-1}$ by $w_i^{(\ell)}$ in (4) and (5), respectively, and also in the computations needed to produce the $\hat{\phi}_i$ in (9). In section 5.4 below, we evaluate the practical performance of the replication variance procedure on NHANES data.

5. Application to NHANES data

5.1 The NHANES design

We evaluate the performance of the local polynomial adjusted estimators on real data. We consider the 2005–2006 release of the National Health and Nutrition Examination Survey (NHANES), which is conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention (NCHS/CDC), of the U.S. Department of Health and Human Services. This survey consists of a stratified, multistage sample of the U.S. civilian non-institutionalized population. A general overview of the sample formation is as follows:

- (i) within each stratum, primary sampling units (PSUs) consisting of counties or grouped smaller counties are selected by sampling with probabilities proportional to a measure of size;
- (ii) from the sampled PSUs, groups of city blocks (segments) containing clusters of households are selected also by sampling with probability proportional to size;
- (iii) in the selected segments, clusters of households are randomly selected with varying selection probabilities to oversample groups of age, ethnic, or income in certain geographic areas; and
- (iv) in the selected households, one or more participants are selected randomly.

The public release of NHANES data has two important aspects. First, to reduce disclosure risks, the stratified, four-stage survey is condensed in a stratified one-stage design, with neither the new stratum variable nor the new PSU variable corresponding to the same variables in the original design. Secondly, the base sampling weights, obtained by reciprocal of the inclusion probabilities of the survey participants, are not released. The weights provided reflect adjustments made to the base weights to account for unit nonresponse, in the interview and exam portions of the survey, and to produce estimates satisfying known population controls.

5.2 The simulation experiment

In order to empirically evaluate the local polynomial estimators as adjustments for nonresponse in complex surveys, we will apply an artificially generated source of unit nonresponse to the public-release NHANES dataset. The nonresponse mechanism will be taken as a smooth function of the age in years of the survey participant (AGE). For this comparison, we chose as study variables four characteristics related to heart diseases, namely the systolic blood pressure (SBP), the diastolic blood pressure (DBP), the indicator of hypertension (HTN) and the indicator of high serum total cholesterol (HTC). All of these were measured on survey participants who were 18 years or older. The systolic and diastolic variables were obtained as the average of the corresponding measurements in a set of up to four readings. Hypertension was defined for individuals having systolic blood pressure of 140 mm Hg or higher or a mean diastolic blood pressure of 90 mm Hg or higher or currently taking medication to lower high blood pressure. High serum total cholesterol was considered when the individual had a total serum cholesterol greater than or equal to 240 mg/dL. The unweighted sample correlations among these and the AGE variable are 0.481 (SBP), 0.118 (DBP), 0.552 (HTN) and 0.060 (HTC), respectively. Hence, it is reasonable to postulate that unit nonresponse related to age is likely to have different effects on survey estimators for these four variables.

The total number of eligible individuals in the NHANES dataset is 4,727. We generated unit nonresponse for the four variables of interest according to two logistic response propensity functions of the auxiliary variable x taken by the age (in years) of the survey participant minus 18. These functions consider a linear and a nonlinear predictor of x as follows

Linear predictor:

$$\phi_I(x) = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1}$$

Nonlinear predictor:

$$\phi_{II}(x) = \{1 + \exp[-(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \cos(\beta_4 x^2/\pi) \sin(\beta_5 x/\pi))]\}^{-1},$$

where the regression coefficients β_0, \dots, β_5 were chosen so that the response propensity functions give an overall nonresponse rate of about 30% when applied to the sample values of x . In both cases, we kept the NHANES sample fixed and generated $B = 1,000$ independent response indicator vectors by Poisson sampling.

The following six nonresponse adjustments were evaluated on these data. Note that in all cases we reported the ratio versions (5) of the estimators, because they were found

to be much more precise than the Horvitz-Thompson versions.

1. True response probabilities: $\hat{\phi}_i = \phi(x_i)$, $i \in s_v$.
2. Logistic regression adjustment: $\hat{\phi}_i$ obtained as the estimated probabilities from a logistic regression of each response vector on x , using a polynomial in x of degree one as the linear predictor.
3. Weighted local polynomial regression of degree k and bandwidth h_v : $\hat{\phi}_i = \hat{\phi}(x_i, k, h_v)$ given by (8), with $i \in s_v$, $k = 0, 1, 2, 3$, $h_v = 0.15, 0.25, 0.50$ and the Epanechnikov kernel function

$$K(t) = (3/4)(1 - t^2)I\{|x| \leq 1\}.$$

4. Unweighted local polynomial regression of degree k and bandwidth h_v : the same as above but not including the sampling weights in (8) to obtain the $\hat{\phi}_i = \hat{\phi}(x_i, k, h_v)$. This might be somewhat easier to compute in practice and should lead to similar results, even if it does not, strictly speaking, follow the pseudo-randomization theory of Section 3.
5. Weighting within cell: within each stratum, respondents and nonrespondents were classified into four classes of age based on the sample quartiles of this variable. This procedure subdivided the sample in a total of 60 cells. Let s_g and s_{rg} denote respectively the set of sampled elements and the set of responding elements in the g^{th} cell. Then, the WC adjustment is defined by taking

$$\hat{\phi}_i = \frac{\sum_{i \in s_{rg}} w_i}{\sum_{i \in s_g} w_i},$$

for all respondents $i \in s_{rg}$.

6. Naive: $\hat{\phi}_i = 1$, $i \in s_v$.

5.3 Bias and robustness against a misspecified response propensity function

When the full sample without artificial nonresponse is used, the Hájek estimated means for the four study variables are respectively SBP = 122.19 mm Hg, DBP = 70.29 mm Hg, HTN = 29.04% and HTC = 15.76%. Table 1 gives the percentage bias relative to those means across response sets obtained for every adjustment procedure in this simulation experiment. For both weighted and unweighted Local Polynomial Regression adjusted estimators, we only display the results for the bandwidth $h_v = 0.25$, but those for other bandwidth values are similar. We instead show the results for different degrees of the local polynomial, so that the effect of moving from local constant to higher order polynomials can be evaluated.

Table 1
Relative biases (%) of nonresponse-adjusted estimators for mean systolic blood pressure (SBP), diastolic blood pressure (DBP), indicator of hypertension (HTN) and indicator of high serum total cholesterol (HTC), based on 1,000 response sets for two propensity functions of the age of the survey participant in NHANES 2005-2006

Type of adjustment	Logistic propensity function (linear predictor)				Logistic propensity function (nonlinear predictor)			
	SBP	DBP	HTN	HTC	SBP	DBP	HTN	HTC
True Response Propensities	0.01	0.01	-0.01	0.04	-0.00	-0.00	0.01	-0.22
Logistic Regression	0.01	0.00	-0.03	0.03	0.47	-1.67	6.49	-6.76
Weighted Local Polynomial Regression:								
Degree 0	0.27	0.34	3.39	2.41	-0.20	-0.39	-1.20	-2.27
Degree 1	0.00	0.04	-0.03	0.20	-0.01	-0.49	0.34	-2.36
Degree 2	0.01	0.01	0.03	0.07	0.03	-0.05	0.51	-0.27
Degree 3	0.01	0.01	-0.02	0.04	-0.03	-0.05	-0.24	-0.44
Unweighted Local Polynomial Regression:								
Degree 0	0.11	0.24	1.34	1.53	-0.17	-0.47	-0.98	-2.70
Degree 1	0.01	0.05	-0.00	0.25	-0.01	-0.57	0.34	-2.69
Degree 2	0.01	0.01	-0.00	0.07	0.01	-0.07	0.26	-0.40
Degree 3	0.00	0.01	-0.06	0.03	-0.03	-0.06	-0.29	-0.48
Weighting Within Cell	0.08	0.08	0.84	0.69	-0.11	-0.07	-0.84	-0.48
Naïve	1.62	0.80	20.49	8.04	-1.30	-1.60	-15.61	-10.77

Among the estimators affected by the generated nonresponse, the worst bias performances are clearly for the unadjusted "Naïve" estimator. As displayed in the last row of Table 1, the biases are higher in the estimation of the prevalence of hypertension and the mean systolic blood pressure, as these are the characteristics of the study variables with higher correlations with the AGE variable, and also for the prevalence of high serum total cholesterol. The biases of the Naïve estimator can be successfully reduced with the true response propensity estimator, any of the local polynomial regression adjusted estimators, the weighting-within cell estimator or with the logistic adjusted estimator, if the model for the propensity function is correctly specified. The best performances in terms of small bias are obtained using the estimator adjusted by the true response propensities, because it is conditionally unbiased for the full sample estimates. The logistic adjustment when it is applied under the correct model, given by the propensity function with a linear predictor, also gives nearly unbiased estimates. For the second propensity function, where the form of the predictor is not well captured by the logistic regression fit of a regression line, this adjustment yields a conditionally biased estimator.

The averages of the local polynomial regression estimates become generally closer to the full sample estimates by increasing the degree of the polynomial fitted, with the largest jump when moving from a local constant to a local linear estimator. Hence, it seems that local polynomial regression is indeed superior to kernel regression in this context. There is very little difference between the weighted and unweighted forms of this adjustment and both procedures have overall smaller conditional biases than the biases of the weighting-within cell estimator, when they are implemented by fitting locally a polynomial of order greater

than zero to estimate the response propensities. The zero degree propensity weighted and unweighted adjusted estimators have smaller biases at smaller bandwidths, as we observed with the bandwidth 0.15, for instance, but smaller bandwidths tend to increase the variance of the estimators. Overall, both weighted and unweighted local polynomial regression adjustments outperform the parametric logistic adjustment when the response model is misspecified. By implementing the local polynomial adjustments with degrees above one, their performances are similar to the one of the logistic adjustment under the correct specification of the response model.

5.4 Variance and variance estimation

Table 2 shows the variance of the adjustment methods considered here across the nonresponse replicates, and we normalized them by the variance for the true response propensity adjustment for clarity. Interestingly, there appears to be an inverse relationship between the magnitude of the relative biases in Table 1 and the variances in this table. In those cases where the relative bias was small (the weighted and unweighted local polynomial regression, the weighting within cell as well as the logistic regression adjustment for the linear propensity function), all the methods appear to result in roughly similar variances. There is a tendency for higher degree local polynomials to be more variable than lower degree ones, and this is particularly noticeable for the nonlinear propensity function, where a clear jump is seen when one moves from degree 1 (local linear) to 2 (local quadratic). Overall, it seems that local linear regression, either weighted or unweighted, offers a good compromise between the bias and the variance of the nonresponse adjustment procedure.

Table 2

Normalized Monte Carlo variances of nonresponse-adjusted estimators for mean systolic blood pressure (SBP), diastolic blood pressure (DBP), indicator of hypertension (HTN) and indicator of high serum total cholesterol (HTC), based on 1,000 response sets for two propensity functions of the age of the survey participant in NHANES 2005-2006

Type of adjustment	Logistic propensity function (linear predictor)				Logistic propensity function (nonlinear predictor)			
	SBP	DBP	HTN	HTC	SBP	DBP	HTN	HTC
True Response Propensities	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Logistic Regression	85.9	92.4	79.5	96.5	63.9	61.5	54.1	52.0
Weighted Local Polynomial Regression:								
Degree 0	74.9	81.2	65.7	92.1	70.3	67.0	67.4	75.0
Degree 1	81.8	89.5	66.2	92.7	73.6	69.8	68.9	76.0
Degree 2	81.3	89.8	65.5	94.0	90.3	81.7	88.0	96.1
Degree 3	82.3	90.2	65.8	93.1	90.1	82.2	87.7	96.2
Unweighted Local Polynomial Regression:								
Degree 0	82.2	85.8	77.6	95.8	71.9	69.2	70.7	74.7
Degree 1	85.6	90.1	79.4	95.7	74.4	71.1	71.2	74.6
Degree 2	86.6	91.3	79.3	96.1	91.8	84.5	91.8	96.8
Degree 3	87.3	91.5	78.5	95.0	91.2	84.7	91.2	96.9
Weighting Within Cell	79.7	89.1	62.1	91.6	82.5	77.0	81.1	92.3
Naive	71.3	58.0	81.7	74.6	48.6	48.7	45.5	45.1

The above simulation results showed the behavior of several nonresponse adjustments in the NHANES setting. We now consider the replication variance estimation approach of Section 4 and evaluate its usefulness as a sample-based measure of uncertainty for the nonresponse-adjusted estimators in the same setting. We implemented (17) with the Jackknife method. Since NHANES does not provide information on the joint sample inclusion probabilities, we could not apply a full Jackknife variance estimator as in, for instance, Berger and Skinner (2005), as a means to account for the selection of units with varying probabilities in the survey. Because of this, we assumed the within-stratum designs in NHANES could be approximated by cluster sampling with replacement and rewrite (17) in the form proposed by Rust (1985),

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{t=1}^T c_t \sum_{j \in s_t} (\hat{\theta}^{(tj)} - \hat{\theta})^2, \quad (18)$$

where s_t denote the set of units in sample from the t^{th} NHANES stratum, $t = 1, 2, \dots, T$, n_t be the number of units selected to s_t , $c_t = (n_t - 1)/n_t$ and $\hat{\theta}^{(tj)}$ is obtained from (5) by replacing the w_i with the replication weights

$$w_{i(j)} = \begin{cases} 0, & \text{for a survey participant} \\ & i \in \text{PSU } j, j \in s_t \\ n_t / (n_t - 1) w_i, & \text{for a survey participant} \\ & i \in \text{PSU } j', j' \in s_t (j' \neq j) \\ w_i, & \text{for a survey participant} \\ & i \notin s_t. \end{cases}$$

These weights were also applied in the estimation of the response propensities for the weighted local polynomial regression procedure adjustment procedure.

The Jackknife variance estimator (18) was applied to each response vector from the two propensity functions, yielding estimates $\hat{v}_{JK}(\hat{\theta}(b))$, $b = 1, 2, \dots, B$, for all adjusted estimators in the experiment. For the sake of comparison, it would be informative to produce estimates of the corresponding variances by the Monte Carlo method. However, as the NHANES sample is fixed, the Monte Carlo variance of the point estimates $\hat{\theta}(b)$ across response vectors estimates only the conditional variance $\text{Var}(\hat{\theta}|s_v)$ with respect to the response model. Since

$$\text{Var}(\hat{\theta}) = \text{Var}(E(\hat{\theta}|s_v)) + E(\text{Var}(\hat{\theta}|s_v)),$$

where the “inner” moments are taken with respect to the response model given the sample and the “outer” moments are with respect to the sampling design, the design variance of $E(\hat{\theta}|s_v)$ needs to be accounted for in order to have a valid estimation target for $\hat{V}_{JK}(\hat{\theta})$. Using the fact that weighted and unweighted local polynomial regression and weighting within cell all produce approximately conditionally unbiased estimators of the full sample estimator, $\bar{y}_{\pi, \text{rat}} = \sum_{i \in s_v} w_i y_i / \sum_{i \in s_v} w_i$, for the two response propensities functions, we decided to use the Jackknife variance estimator of $\bar{y}_{\pi, \text{rat}}$ as a “proxy” for $\text{Var}(E(\hat{\theta}|s_v))$. Hence, our “comparison variance” will be defined as

$$\hat{v}_C(\hat{\theta}) = \hat{v}_{JK}(\bar{y}_{\pi, \text{rat}}) + \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}(b) - \hat{\theta})^2.$$

Using $\hat{v}_c(\hat{\theta})$ instead of the true variance will tend to underestimate any bias issues associated with the use of the jackknife variance estimator for the full sample estimator. However, it will show how well the replication procedure manages to capture the nonresponse variability.

Table 3 gives relative biases of the Jackknife variance estimators obtained in this experiment. The results show that the jackknife variance estimator performs reasonably well for both nonresponse mechanisms and all estimators considered. The weighted local polynomial regression adjusted procedure appears to yield estimated variances in greater consonance with the comparison variance than when the procedure is implemented by its unweighted version. The results for the nonlinear predictor function exhibit more bias than those for the linear predictor, with more pronounced positive and negative biases present for the former for all the variables. As discussed in Da Silva and Opsomer (2006), replication methods for nonresponse-adjusted estimators often ignore a component of the total variance, which includes the effect of both sampling and the response mechanism. We therefore conjecture that the different bias behaviors exhibited for the different variables could be due to this missing variance component.

6. Concluding remarks

In this article, we studied properties of nonparametric propensity weighting as an adjustment procedure for survey nonresponse. The local polynomial regression technique is seen to offer a flexible way of constructing new survey adjustments for nonresponse. The results in the article extend those in Da Silva and Opsomer (2006) by allowing

the use of local polynomials of arbitrary degree, which offers both theoretical and practical advantages over zero-degree kernel regression.

In addition to its good theoretical properties, the estimator was shown in the simulation experiment to be competitive with an estimator based on a correctly specified parametric model in terms of bias and variance, while protecting against a potentially misspecified model. The weighting-cell estimator is similarly robust against model misspecification, but a particular advantage of nonparametric regression methods over weighting cell approaches is the connection to broad classes of modeling techniques available in the non-survey literature. Extensions of the methodology we described here to semiparametric and (generalized) additive models (Hastie and Tibshirani 1986) are readily formulated and should work well in a wide range of potential response model scenarios, including situations with multiple covariates that are both categorical and continuous. A detailed discussion of these extensions is beyond the scope of the current paper, however.

In Section 5, we applied the nonparametric nonresponse adjustment to NHANES data by modeling the response probability as a smooth function of the age of the respondents, and weighting the data by the inverses of the estimated response probabilities. The same approach can be used in other survey datasets whenever continuous covariates related to the response probability are available for all elements in the original sample. This provides a viable alternative to the commonly used weighting-within-cell approach for situations in which cells are constructed by “binning” one or several continuous variables.

Table 3
Relative biases (%) of the Jackknife variance estimators of estimators of the mean systolic blood pressure (SBP), diastolic blood pressure (DBP), indicator of hypertension (HTN) and indicator of high serum total cholesterol (HTC), based on 1,000 response sets for two propensity functions of the age of the survey participant in NHANES 2005-2006

Type of adjustment	Logistic propensity function (linear predictor)				Logistic propensity function (nonlinear predictor)			
	SBP	DBP	HTN	HTC	SBP	DBP	HTN	HTC
True Response Propensities	0.55	-0.47	-0.06	0.16	0.92	-0.26	-1.03	-2.76
Weighted Local Polynomial Regression:								
Degree 0	-0.66	2.33	2.74	4.44	1.63	-2.27	-5.12	-9.44
Degree 1	-0.31	-1.03	0.31	1.87	5.27	4.03	2.60	-9.95
Degree 2	-0.14	-0.76	0.41	0.49	0.25	0.65	-2.60	-3.60
Degree 3	-0.27	-1.03	0.39	0.48	0.19	0.45	-2.19	-3.02
Unweighted Local Polynomial Regression:								
Degree 0	2.00	2.77	3.57	5.56	5.73	0.31	1.83	-10.22
Degree 1	2.02	1.06	2.63	2.61	7.46	5.57	4.33	-10.43
Degree 2	2.26	1.07	2.88	1.36	4.16	3.81	1.62	-2.94
Degree 3	2.21	1.01	2.94	1.46	3.45	3.65	0.96	-2.63
Weighting Within Cell	-1.15	1.70	-0.47	5.16	2.69	-6.91	3.06	-5.88

There are still a number of open issues that need to be further investigated with respect to implementation of the method in actual surveys, whether in the univariate case described in detail here or in the various model extensions just mentioned. An important practical issue is the selection of estimator settings such as the degree of the local polynomial and the bandwidth. As noted in the nonparametric literature (e.g., Fan and Gijbels 1996, page 77) and also confirmed in the simulations, higher degree polynomials reduce the bias but increase the variance, so that polynomials of degree $k = 1$ or 2 are generally recommended as a good compromise. More critical is the choice of bandwidth parameter. In our simulations, the results were only modestly sensitive to the choice of bandwidth within a “reasonable” range of values, i.e., ones ensuring that the number of observations used for estimating $\phi(x)$ at any x does not become too small (see discussion at the end of Section 2), or that is so large that the fit cannot capture changes in $\phi(\cdot)$ over the range of x . As a rule of thumb, we would recommend considering values for h that are within 20% and 50% of the range of x as a good place to start, and making a final determination by looking at both model diagnostics for the model fit $\hat{\phi}(x)$ and weight diagnostics for the adjusted survey weights $(\pi_i \hat{\phi}_i)^{-1}$, similarly as would be done when constructing cell-based weights.

Acknowledgements

We thank the Associate Editor and two referees for their useful comments. The first author was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, under the grant Projeto Universal 480518/2004-1.

Appendix

A.1 Assumptions

We now state the assumptions needed to derive our main results. A detailed discussion of these assumptions is provided in Da Silva and Opsomer (2008). Consider the asymptotic framework of Section 3. Let $\mathbf{I}_v = (I_1, I_2, \dots, I_{N_v})'$ be the sample inclusion indicator vector for the v^{th} population. Suppressing the v for ease of notation, let $\pi_i = \Pr(I_i = 1)$, and let

$$\Delta_{j_1, \dots, j_k} \equiv E_d \left(\prod_{\ell=1}^k (I_{j_\ell} - \pi_{j_\ell}) \right) \quad (19)$$

denote higher moments for the sample inclusion indicators $I_{j_1}, I_{j_2}, \dots, I_{j_k}$ with respect to the sampling design. We

assume that there are positive constants $\lambda_1, \lambda_2, \dots, \lambda_6$ such that:

$$(A1) \lambda_1 < N_v n_v^{-1} \pi_i < \lambda_2 < \infty, \forall i \in U_v;$$

$$(A2) N_v^{-1} n_v \rightarrow \pi, \text{ for some } 0 < \pi < 1, \text{ as } v \rightarrow \infty;$$

$$(A3) \text{ For distinct } j_1, j_2, \dots, j_k \in U_v, \text{ where } k = 2, 3, \dots, 8,$$

$$|\Delta_{j_1, \dots, j_k}| \leq \begin{cases} \left[\prod_{\ell=1}^k (N - \ell + 1) \right]^{-1} n_v^{\frac{k}{2}} \lambda_3 & \text{if } k \text{ is even,} \\ \left[\prod_{\ell=1}^k (N - \ell + 1) \right]^{-1} n_v^{\frac{k-1}{2}} \lambda_4 & \text{if } k \text{ is odd} \end{cases}$$

$$(A4) \lim_{v \rightarrow \infty} N_v^{-1} \sum_{i \in U_v} y_i = \mu \in (-\infty, \infty) \text{ and } N_v^{-1} \sum_{i \in U_v} |y_i|^4 \leq \lambda_5, \text{ for all } v \geq 1.$$

Let $\mathbf{R}_v = (R_1, R_2, \dots, R_{N_v})'$ denote the response indicator vector for the v -th population. In addition to the assumptions on the sampling design and the population distribution of the variable Y , we will also need the following assumptions on the response mechanism:

$$(B1) R_1, R_2, \dots, R_{N_v} \text{ are independent random variables;}$$

$$(B2) \Pr\{R_i = 1 \mid \mathbf{I}_v, \mathbf{y}_v, \mathbf{x}_v\} =$$

$$\Pr\{R_i = 1 \mid \mathbf{x}_v\} \equiv \phi_i, \forall i \in U_v;$$

$$(B3) \phi_i = \phi(x_i), \forall i \in U_v, \text{ where } \phi(\cdot) \text{ is a } (k+2)^{\text{th}} \text{ continuously differentiable function with } \lambda_6 < \phi(\cdot) \leq 1. \text{ The first derivative } \phi'(\cdot) \text{ has a finite number of sign changes.}$$

Regarding the distribution of the x_i and the kernel estimator, we assume that:

$$(C1) \text{ For all } v \geq 1, x_1, x_2, \dots, x_{N_v} \text{ are realizations of random variables } X_1, X_2, \dots, X_{N_v} \text{ independent and identically distributed with distribution } F_X(x) = \int_{-\infty}^x f_X(t) dt, \text{ where } f_X(\cdot) \text{ is a continuous and positive probability density function on a compact set } [a_X, b_X];$$

$$(C2) \text{ The kernel function } K(\cdot) \text{ is a bounded and continuous probability density, which is symmetric around zero and supported on } [-1, 1];$$

$$(C3) \int_{-1}^1 |z|^{k+4} K(z) dz < \infty;$$

$$(C4) \text{ For all } v \geq 1, \{h_v\} \text{ is a sequence of bandwidths satisfying } 0 < h_v \leq 1, h_v \rightarrow 0, n_v h_v^2 \rightarrow \infty \text{ and } N_v h_v / \log N_v \rightarrow \infty, \text{ as } v \rightarrow \infty;$$

$$(C5) \text{ The first derivative } f'_X(\cdot) \text{ is continuously differentiable and contains a finite number of sign changes on } \text{supp}(f_X). \text{ The first derivative } K'(\cdot) \text{ has a finite number of sign changes on } \text{supp}(K);$$

(C6) The matrix $N_v \mathbf{T}_i^{-1}$ is non-singular for all $i \in U_v$ and all $v \geq 1$.

A.2 Technical derivations

Complete proofs are in Da Silva and Opsomer (2008). The proof of Theorem 1 relies on bounding the moments of the difference $\bar{y}_{\pi\psi v} - \bar{y}_{\pi\phi v}$ under the combined design and response model probability mechanism, followed by deriving the rates of convergence for the bias and variance of the linearized estimator $\bar{y}_{\pi\psi v}$. This is done in a series of six lemmas, which are stated here without proof. The proof of Theorem 2 is based on the result of Theorem 1, followed by an additional linearization of the ratio form.

For notational simplicity in what follows, we suppressed the fact that the results are conditional on the sequences $\mathbf{x}_v = (x_1, \dots, x_{N_v})$ in the populations U_v . However, the results in these lemmas are shown to hold with probability one over these sequences in Da Silva and Opsomer (2008), as was also done in Da Silva and Opsomer (2006). Hence, the results can be interpreted to hold for all population sequences, except on a set of probability 0 with respect to the distribution of the \mathbf{x}_v .

Lemma 1. Assume that assumptions (C1)-(C5) hold. Consider $\mu_\ell(K, x) = \int_{D_{x,h_v}} z^\ell K(z) dz$, where $D_{x,h_v} = \{t: (x + ht) \in \text{supp}(f_X)\} \cap \text{supp}(K)$. Then, for all $\ell = 0, 1, \dots, k+2$,

$$\sup_{x \in \text{supp}(f_X)} \left| \frac{1}{N_v h_v} \sum_{j \in U_v} K\left(\frac{X_j - x}{h_v}\right) (X_j - x)^\ell - E_v(x, \ell) \right| \xrightarrow{v \rightarrow \infty} 0,$$

where

$$E_v(x, \ell) = f_X(x) \mu_\ell(K, x) h_v^\ell + f'_X(x) \mu_{\ell+1}(K, x) h_v^{\ell+1} + o(h_v^{\ell+1}).$$

Lemma 2. Assume that assumptions (C1)-(C5) hold. Consider the population fit $\tilde{\phi}_i = \tilde{\phi}(x_i, k, h_v)$, $i \in U_v$, defined in (10). Hence, for all $i \in U_v$, there exists positive bounded terms $c_1(x_i)$, $c_2(x_i)$ and $c_3(x_i)$, such that if x_i in an interior point of $\text{supp}(f_X)$

$$\tilde{\phi}_i - \phi(x_i) = \begin{cases} c_1(x_i) h_v^{k+2} + o(h_v^{k+2}) & k \text{ is even} \\ c_2(x_i) h_v^{k+1} + o(h_v^{k+1}) & k \text{ is odd} \end{cases}$$

and if x_i in a boundary point of $\text{supp}(f_X)$

$$\tilde{\phi}_i - \phi(x_i) = c_3(x_i) h_v^{k+1} + o(h_v^{k+1}),$$

where all the smaller order terms hold uniformly in $i \in U_v$.

Lemma 3. Assume that assumptions (C1) and (C4) hold. Then,

i) For $p \in [0, \infty)$ fixed,

$$\limsup_{v \rightarrow \infty} \left(\frac{1}{N_v h_v} \sum_{j \in U_v} I_{\{x - h_v \leq x_j \leq x + h_v\}} \right)^p < \infty,$$

uniformly in x ;

$$\text{ii) } \limsup_{v \rightarrow \infty} \frac{1}{2N_v h_v} \sum_{j \in U_v} I_{\{x_j \in [0, h_v] \cup (1 - h_v, 1]\}} < \infty;$$

$$\text{iii) } \limsup_{v \rightarrow \infty} \frac{1}{N_v} \sum_{j \in U_v} I_{\{x_j \in (h_v, 1 - h_v)\}} < \infty.$$

iv) there exists v^* , independent of x , such that whenever $v \geq v^*$,

$$\sum_{j \in U_v} I_{\{|x_j - x| \leq h_v\}} \geq k + 1;$$

Lemma 4. Suppose the assumptions of Theorem 1 hold. Consider the matrices $\hat{\mathbf{T}}_{si} = \{\hat{T}_{si, pq}\}$ and $\mathbf{T}_i = \{T_{si, pq}\}$ and the vectors $\hat{\mathbf{t}}_{si} = \{\hat{t}_{si, p}\}$, $\mathbf{t}_i = \{t_{i, p}\}$ and $\mathbf{B}_i = \{B_{i, p}\}$ given in (7) and (10). Then,

- i) the $N_v^{-1} T_{i, pq}$ and $N_v^{-1} t_{i, p}$ are uniformly bounded in $i \in U_v$, for all $p, q = 1, \dots, k+1$;
- ii) the $\hat{T}_{si, pq}$ and $\hat{t}_{si, p}$ satisfy

$$\begin{aligned} \max_{1 \leq p, q \leq k+1} E \left(\frac{\hat{T}_{si, pq} - T_{i, pq}}{N_v} \right)^8 &= O \left(\frac{1}{n_v^4 h_v^4} \right) \text{ and} \\ \max_{1 \leq p \leq k+1} E \left(\frac{\hat{t}_{si, p} - t_{i, p}}{N_v} \right)^8 &= O \left(\frac{1}{n_v^4 h_v^4} \right), \end{aligned}$$

uniformly in $i \in U_v$;

- iii) the random variable $e'_i \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si} \mathbf{B}_i)$ satisfies

$$\max_{i \in U_v} E (e'_i \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si} \mathbf{B}_i) I_i R_i) = O \left(\frac{1}{n_v h_v} \right) \quad (20)$$

and

$$\max_{i \in U_v} E (e'_i \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si} \mathbf{B}_i))^4 = O \left(\frac{1}{n_v^2 h_v^2} \right). \quad (21)$$

Lemma 5. Suppose the assumptions of Theorem 1 hold. Then, for all $v \geq 1$

- i) the reciprocal of $\tilde{\phi}_i$ is uniformly bounded in $i \in U_v$;
- ii) the partial derivatives of $\tilde{\phi}_i^{-1}$ of orders one up to four, when evaluated at $\hat{\mathbf{T}}_{si} = \mathbf{T}_i$, $\hat{\mathbf{t}}_{si} = \mathbf{t}_i$, $\delta_1 = 0$ and $\delta_2 = 0$, are uniformly bounded in $i \in U_v$;
- iii) $E(\tilde{\phi}_i^{-4})$ is uniformly bounded in $i \in U_v$;
- iv) the reciprocal of $\tilde{\phi}_i$ satisfies

$$\hat{\phi}_i^{-1} = \bar{\phi}_i^{-1} - \bar{\phi}_i^{-2} \mathbf{e}_i' \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_i \mathbf{B}_i) + \varepsilon_{iv} + O\left(\frac{1}{N_v^2 h_v^2}\right), \quad (22)$$

uniformly in $i \in U_v$, where the ε_{iv} are random variables such that

$$\max_{i \in U_v} E(\varepsilon_{iv}^2) = O\left(\frac{1}{n_v^2 h_v^2}\right).$$

Lemma 6. Suppose the assumptions of Theorem 1 hold. Define the random variables $\bar{y}_{\pi\hat{\phi}_v}$, $\bar{d}_{\pi\hat{\phi}_v}$ and $\bar{\varepsilon}_{\pi\hat{\phi}_v}$ as

$$(\bar{y}_{\pi\hat{\phi}_v}, \bar{d}_{\pi\hat{\phi}_v}, \bar{\varepsilon}_{\pi\hat{\phi}_v})' = \frac{1}{N_v} \sum_{i \in S_v} \pi_i^{-1} \bar{\phi}_i^{-1} (1, \bar{\phi}_i^{-1} \mathbf{e}_i' \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_i \mathbf{B}_i), \varepsilon_{iv}) y_i R_i.$$

Then,

$$E(\bar{y}_{\pi\hat{\phi}_v} - \bar{y}_{N_v}) = \begin{cases} O(h_v^{k+(3/2)}) & k \text{ even}, \\ O(h_v^{k+1}) & k \text{ odd}, \end{cases} \quad (23)$$

$$\text{Var}(\bar{y}_{\pi\hat{\phi}_v}) = O\left(\frac{1}{n_v}\right), \quad (24)$$

$$(E[\bar{d}_{\pi\hat{\phi}_v}], E[\bar{d}_{\pi\hat{\phi}_v} \hat{A}])' = O\left(\frac{1}{n_v h_v}\right) \quad (25)$$

and

$$E(\bar{\varepsilon}_{\pi\hat{\phi}_v}^2) = O\left(\frac{1}{n_v^2 h_v^2}\right). \quad (26)$$

References

- Alho, J.M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika*, 617-624.
- Berger, Y.G., and Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 67(1), 79-89.
- Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 1026-1053.
- Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete data in sample surveys: Theory and bibliographies*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin). Academic Press, New York: London, 3, 143-160.
- Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics*, 4, 563-579.
- Da Silva, D.N., and Opsomer, J.D. (2008). Theoretical properties of propensity weighting for survey nonresponse through local polynomial regression. Technical Report #2008/6, Department of Statistics, Colorado State University.
- David, M.H., Little, R., Samuël, M. and Triest, R. (1983). Imputation models based on the propensity to respond. In *ASA Proceedings of the Business and Economic Statistics Section*, 168-173.
- Eckholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 325-337.
- Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.
- Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *ASA Proceedings of the Social Statistics Section*, 197-202.
- Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron*, 4, 185-200.
- Groves, R., Dillman, D., Eltinge, J. and Little, R.J.A. (2002). *Survey Nonresponse*. New York: John Wiley & Sons, Inc.
- Hastie, T.J., and Tibshirani, R.J. (1986). Generalized additive models. *Statistical Science*, 297-318.
- Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. In *ASA Proceedings of the Section on Survey Research Methods*, 637-642.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 89-96.
- Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 4, 501-514.
- Laaksonen, S. (2006). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications*, 2, 95-100.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.
- Nargundkar, M., and Joshi, G.B. (1975). Non-response in sample surveys. In *40th Session of the ISI, Warsaw 1975, Contributed papers*, 626-628.
- Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustments for unit non-response. In *Incomplete data in sample surveys*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin), *Theory and bibliographies*, Academic Press, New York: London, 2, 143-184.
- Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41-55.
- Rust, K. (1985). Variance estimation for complex estimators in sample survey. *Journal of Official Statistics*, 381-397.
- Wand, M.P., and Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design

Jan van den Brakel and Sabine Krieg¹

Abstract

In this paper a multivariate structural time series model is described that accounts for the panel design of the Dutch Labour Force Survey and is applied to estimate monthly unemployment rates. Compared to the generalized regression estimator, this approach results in a substantial increase of the accuracy due to a reduction of the standard error and the explicit modelling of the bias between the subsequent waves.

Key Words: Small area estimation; Rotation group bias; Survey errors.

1. Introduction

The Dutch Labour Force Survey (LFS) is based on a rotating panel design. Each month a sample of addresses is drawn and data are collected by means of computer assisted personal interviewing of the residing households. The sampled households are re-interviewed by telephone four times at quarterly intervals. The estimation procedure of this survey is based on the generalized regression (GREG) estimator, developed by Särndal, Swensson and Wretman (1992).

Due to the following properties, GREG estimators are very attractive to produce official releases in a regular production environment and are therefore widely applied by national statistical institutes. First, GREG estimators are approximately design-unbiased, which provides a form of robustness in the case of large sample sizes. These estimators are derived from a linear regression model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables for which the totals in the finite target population are known. If this linear regression model explains the variation of the target variable reasonably well, then this might reduce the design variance as well as the bias due to selective nonresponse, Särndal and Swensson (1987), Bethlehem (1988), and Särndal and Lundström (2005). Model misspecification, on the other hand, might result in an increase of the design variance but the point estimates remain approximately design unbiased. Second, GREG estimators are often used to produce one set of weights for the estimation of all target parameters of a multi-purpose sample survey. This is not only convenient but also enforces consistency between the marginal totals of different publication tables.

There are two major problems with the rotating panel design of the LFS and the way that the GREG estimator is applied in the estimation procedure. First, there are

substantial systematic differences between the subsequent waves of the panel due to mode- and panel effects. This is a well-known problem for rotating panel designs, and is in the literature referred to as rotation group bias (RGB), see Bailer (1975). In the LFS, the level of the unemployment rate in the subsequent waves is substantially smaller compared to the first wave. There are also systematic differences between the seasonal effects of the subsequent waves.

A second problem is that the monthly sample size of the LFS is too small to rely on the GREG estimator to produce official statistics about the monthly employment and unemployment. GREG estimators have a relatively large design variance in the case of small sample sizes. Therefore, in the LFS, each month the samples observed in the preceding three months are used to estimate quarterly figures about the labour market situation. The major drawback of this approach is that the real monthly seasonal pattern in the unemployment rate is smoothed out. Also structural changes in unemployment appear delayed in the series of quarterly figures.

Since the monthly sample size is too small to apply design-based or direct survey estimators, model-based estimation procedures might be used to produce sufficiently reliable statistics. In the case of continuously conducted surveys, a structural time series model can be applied to use information from preceding samples to improve the accuracy of the estimates. This model can be extended to account for the RGB and the autocorrelation (AC) between the different panels of the LFS. This approach makes efficient use of the rotating panel design of the LFS in estimating monthly figures about the labour market, and is originally proposed by Pfeffermann (1991) and Pfeffermann, Feder and Signorelli (1998). These techniques are applied in this paper to estimate the monthly unemployment rate of the LFS. Other references to authors that apply

1. Jan van den Brakel and Sabine Krieg, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401CZ Heerlen, The Netherlands.
E-mail: jbrl@cbs.nl and skrg@cbs.nl.

time series models to develop estimates for periodic surveys are Scott and Smith (1974), Scott, Smith and Jones (1977), Tam (1987), Binder and Dick (1989, 1990), Bell and Hillmer (1990), Tiller (1992), Rao and Yu (1994), Pfeffermann and Burck (1990), Pfeffermann and Rubin-Bleuer (1993), Pfeffermann and Tiller (2006), Harvey and Chung (2000), and Feder (2001).

Composite estimators can be considered as an alternative to time series models. They are developed under the traditional design-based approach, to use information observed in previous periods from periodic surveys with a rotating panel design, to improve the precision of level and change estimates. Some key references to composite estimators are Hansen, Hurwitz and Meadow (1953), Rao and Graham (1964), Gurney and Daly (1965), Cantwell (1990), Singh (1996), Gambino, Kennedy and Singh (2001), Singh, Kennedy and Wu (2001) and Fuller and Rao (2001).

In Section 2, the survey design of the LFS is summarised. A structural time series model that accounts for the rotating panel design of the LFS is developed in Sections 3 and 4. The results are detailed in Section 5. Some general remarks are made in Section 6.

2. The dutch Labour Force Survey

2.1 Sample design

The objective of the Dutch LFS is to provide reliable information about the labour market. Each month a sample of addresses is selected from which households are identified that can be regarded as the ultimate sampling units. The target population of the LFS consists of the non-institutionalised population aged 15 years and over residing in the Netherlands. The sampling frame is a list of all known occupied addresses in the Netherlands, which is derived from the municipal basic registration of population data. The LFS is based on a stratified two-stage cluster design of addresses. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. All households residing at an address, up to a maximum of three, are included in the sample (in the Netherlands, there is generally one household per address). Since most target parameters of the LFS concern people aged 15 through 64 years, addresses with only persons aged 65 years and over are undersampled.

In October 1999, the LFS changed from a continuous survey to a rotating panel design. In the first wave, data are collected by means of computer assisted personal interviewing (CAPI). For all members of the selected households, demographic variables are observed. For the target variables only persons aged 15 years and over are interviewed. When a household member cannot be contacted,

proxy interviewing is allowed by members of the same household. Households, in which one or more of the selected persons do not respond for themselves or in a proxy interview, are treated as nonresponding households. The respondents aged 15 through 64 years are re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing (CATI). During these re-interviews a condensed questionnaire is applied to establish changes in the labour market position of the respondents. Proxy interviewing is also allowed during these re-interviews. The monthly gross sample size averaged about 8,000 addresses when the LFS first changed to a rotating panel design. The monthly sample size gradually declined to about 6,500 addresses in 2008. During this period about 65% completely responding households are obtained.

2.2 Rotation group bias

The rotating panel design, described in Section 2.1, results in systematic differences between the estimates of the unemployment rate of the successive waves in one time period. In the literature, this phenomenon is known as RGB, see *e.g.*, Bailer (1975), Kumar and Lee (1983) and Pfeffermann (1991). The RGB in the LFS results in a systematic underestimation of the level of the unemployment rate in the CATI waves but also in systematic differences between the seasonal patterns. The RGB is a consequence of the following strongly confounded factors:

- Selective nonresponse between the subsequent waves, *i.e.*, panel attrition.
- Systematic differences between the populations that are reached with the CAPI and CATI modes. It is anticipated that these differences are relatively small, since telephone numbers are asked during the first interview. As a result, secret numbers and cell-phone numbers are also called.
- Mode-effects, *i.e.*, systematic differences in the data due to the fact that the interviews are conducted by telephone instead of face to face. Under the CAPI mode the interview speed is lower, respondents are more engaged with the interview and are more likely to exert the required cognitive effort to answer questions carefully. Also less socially desirable answers are obtained under the CAPI mode due to the personal contact with the interviewer. As a result, less measurement errors are expected under the CAPI mode (Holbrook, Green and Krosnick 2003, and Roberts 2007). Van den Brakel (2008) describes an experiment where the CAPI and CATI data collection modes are compared in the first wave of the LFS. It follows that the estimated unemployment rate is significantly smaller under the CATI mode.

- The fraction of proxy interviews is larger under the CATI mode (Van den Brakel 2008). This might result in an increased amount of measurement errors.
- Effects due to differences between the CAPI questionnaire and the CATI questionnaire. The CATI questionnaire is a strongly condensed version of the CAPI questionnaire since the re-interviews focus on changes in the labour market position of the respondents.
- Panel effects, *i.e.*, systematic changes in the behaviour of the respondents in the panel. For example, questions about activities to find a job in the first wave might increase the search activities of the unemployed respondents in the panel. Respondents might also adjust their answers in the subsequent waves systematically, since they learn how to keep the routing through the questionnaire as short as possible.

It is assumed that the estimates based on the first wave are the most reliable, since CAPI generally results in a higher data quality and the first wave does not suffer from the panel effects mentioned above. In order to minimize the effects of the RGB, the second, third, fourth and fifth waves are currently calibrated to the first wave as will be described in Section 2.3.

2.3 Regular estimation procedure

Target parameters about the employment and unemployment are defined as population totals or as ratios of two population totals. The unemployment rate, which is investigated in this paper, is defined as the ratio of the total unemployment to the total labour force. This population parameter is estimated as the ratio of the GREG estimate for the total unemployed labour force to the estimated total labour force. Each month estimates about the employment and unemployment for the preceding three months are published.

In an attempt to correct for the RGB, a rather laborious weighting procedure is used in the regular estimation procedure. The most important steps are summarized here. First, the inclusion probabilities are derived, which reflect the sampling design described above as well as the different response rates between geographical regions. Subsequently, the inclusion weights of each CATI wave are calibrated with the GREG estimator to the labour force status observed in the first wave. In the next step, the calibrated weights of the CATI waves and the inclusion weights of the CAPI wave are used as the design or starting weights of the GREG estimator, using a weighting scheme that is based on a combination of different socio-demographic classifications. The integrated method for weighting persons and families of Lemaître and Dufour (1987) is applied to obtain equal weights for persons belonging to the same household. Finally, a bounding algorithm proposed by Huang and

Fuller (1978) is applied to avoid negative weights. This estimation procedure is conducted with the software package Bascula, Nieuwenbroek and Boonstra (2002).

Since this weighting procedure hardly corrects for the RGB, an additional rigid correction is applied. For the most important parameters the ratio between the estimates based on CAPI only and the estimates based on all waves is computed using the data of 12 preceding quarters. Estimates for the preceding three months are multiplied by this ratio to correct for RGB.

2.4 Monthly GREG estimates based on monthly data

In Section 3, a structural time series model is developed to estimate the monthly unemployment rate. The input data for this time series model are the GREG estimates for the monthly unemployment rate using the monthly sample data of the separate waves. Let θ_t denote the true but unknown unemployment rate for month t . Now Y_t^{t-j} denotes the GREG estimate of the unemployment rate of month t , based on the sample which entered the panel in month $t - j$. For the period of January 2001 until December 2008 each month five independent GREG estimates for the same parameter θ_t are produced, using the five separate waves that are observed each month, *i.e.*, Y_t^{t-j} for $j = 0, 3, 6, 9, 12$. These estimates are defined as

$$Y_t^{t-j} = \frac{y_{y,t}^{t-j}}{z_{z,t}^{t-j}}, \quad (2.1)$$

with $y_{y,t}^{t-j}$ and $z_{z,t}^{t-j}$ the GREG estimates for the unemployed labour force and the labour force at time t , based on the sample that entered the panel at $t - j$.

The separate monthly waves are weighted with a reduced version of the weighting scheme that is applied in the regular weighting procedure for the quarterly figures. The estimates based on the CATI data are not adjusted to correct for RGB, since a multivariate time series model is applied to correct for this bias.

The variance of (2.1) can be estimated with

$$\text{var}(Y_t^{t-j}) = \frac{1}{(z_{z,t}^{t-j})^2} \sum_{h=1}^H \frac{n_{h,t}^{t-j}}{n_{h,t}^{t-j} - 1} \left(\sum_{k=1}^{n_{h,t}^{t-j}} (w_k e_{k,t}^{t-j})^2 - \frac{1}{n_{h,t}^{t-j}} \left(\sum_{k=1}^{n_{h,t}^{t-j}} w_k e_{k,t}^{t-j} \right)^2 \right), \quad (2.2)$$

with

$$e_{k,t}^{t-j} = \sum_{l=1}^{m_l} (y_{kl,t}^{t-j} - \mathbf{x}_{kl}^T \mathbf{b}_y) - Y_t^{t-j} (z_{kl,t}^{t-j} - \mathbf{x}_{kl}^T \mathbf{b}_z).$$

Here $y_{kl,t}^{t-j}$ is a binary variable taking value one if the l^{th} person belonging to the k^{th} household that entered the sample at time $t - j$ belongs to the unemployed labour force at time t and zero otherwise, $z_{kl,t}^{t-j}$ a binary variable taking value one if the l^{th} person of the k^{th} household

belongs to the labour force at time t and zero otherwise, \mathbf{x}_{kl} a vector with the auxiliary information of the l^{th} person belonging to the k^{th} household used in the weighting scheme of the GREG estimator, \mathbf{b}_y and \mathbf{b}_z the regression coefficient of the regression function of $y_{kl,t}^{t-j}$ respectively $z_{kl,t}^{t-j}$ on \mathbf{x}_{kl} , w_k the regression weight of household k , $n_{k,t}^{t-j}$ the number of completely responding households of stratum $h = 1, \dots, H$, at time t of the sample that entered the panel at $t - j$, and m_k the number of persons aged 15 years and over belonging to the k^{th} household. Recall from Section 2.3 that persons belonging to the same household have equal weights due to the application of the integrated method for weighting persons and families of Lemaître and Dufour (1987). Formula (2.2) is the variance estimation procedure implemented in Bascula to approximate the variance of the ratio of two GREG estimators.

The estimates for the monthly unemployment rate obtained with the structural time series approach will be compared in Section 5.3 with monthly estimates based on the GREG estimator using the data observed in the five waves. For this comparison a slightly simplified version of the procedure described in Section 2.3 is applied to combine the data observed in the different waves to obtain monthly GREG estimates. First, a GREG estimate Y_t is computed using the data observed in the five waves using the same weighting procedure used in the regular production process to estimate quarterly figures, see Section 2.3. The weighting scheme is slightly simplified because less data are available. Subsequently a correction factor based on the preceding three years is computed as:

$$c_t = \frac{\sum_{j=0}^{35} Y_{t-j}^{t-j}}{\sum_{j=0}^{35} Y_{t-j}}. \quad (2.3)$$

Finally, the corrected estimate is computed:

$$Y_t^c = c_t Y_t. \quad (2.4)$$

Because the series start at January 2001, c_t can be computed from December 2003. To get a corrected GREG estimate for all months, $c_{\text{December2003}}$ is used in formula (2.4) for the periods preceding December 2003. The variance of (2.4) is approximated by $\text{var}(Y_t^c) = c_t^2 \text{var}(Y_t)$, where $\text{var}(Y_t)$ is computed with formula (2.2), using the data of all waves accordingly.

3. Time series model

Direct estimators, like the Horvitz-Thompson estimator or the GREG estimator, assume that the monthly unemployment rate θ_t is a fixed but unknown population parameter. Under this design-based approach, an estimator

for θ_t for cross-sectional surveys only uses the data observed at time t . Data from the past are only used in the case of partially overlapping samples in a panel design, but not in the case of repeatedly conducted cross-sectional designs. Scott and Smith (1974) proposed to consider the population parameter θ_t as a realization of a stochastic process that can be described with a time series model. Under this assumption, data observed in preceding periods $t - 1, t - 2, \dots$, can be used to improve the estimator for θ_t , even in the case of non-overlapping sample surveys.

Recall from Section 2.4 that Y_t^{t-j} denotes the GREG estimator for θ_t based on the panel observed at time t , which entered the survey for the first time at $t - j$. Due to the applied rotation pattern, each month a vector $\mathbf{Y}_t = (Y_t^t Y_t^{t-3} Y_t^{t-6} Y_t^{t-9} Y_t^{t-12})^T$ is observed. According to Pfeffermann (1991), this vector can be modelled as

$$\mathbf{Y}_t = \mathbf{1}_5 \theta_t + \boldsymbol{\lambda}_t + \boldsymbol{\gamma}_t + \mathbf{e}_t, \quad (3.1)$$

with $\mathbf{1}_5$ a five dimensional vector with each element equal to one, $\boldsymbol{\lambda}_t = (\lambda_t^0 \lambda_t^3 \lambda_t^6 \lambda_t^9 \lambda_t^{12})^T$ and $\boldsymbol{\gamma}_t = (\gamma_t^0 \gamma_t^3 \gamma_t^6 \gamma_t^9 \gamma_t^{12})^T$ vectors with time dependent components that account for the RGB in the trend and the RGB in the seasonal components respectively, and $\mathbf{e}_t = (e_t^0 e_t^3 e_t^6 e_t^9 e_t^{12})^T$ the corresponding survey errors for each panel estimate. Time series models for the different components in (3.1), i.e., the population parameter θ_t , the RGB for the trend $\boldsymbol{\lambda}_t$, the RGB for the seasonal patterns $\boldsymbol{\gamma}_t$, and the survey errors \mathbf{e}_t , are developed in Sections 3.1 through 3.3.

3.1 Time series model for the population parameter

With a structural time series model, the population parameter θ_t in (3.1) can be decomposed in a trend component, a seasonal component, and an irregular component, i.e.:

$$\theta_t = L_t + S_t + \varepsilon_t, \quad (3.2)$$

where L_t denotes a stochastic trend component, S_t a stochastic seasonal component, and ε_t the irregular component. For the stochastic trend component the so-called local linear trend model is used, which is defined by the following set of equations:

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1} + \eta_{L,t}, \\ R_t &= R_{t-1} + \eta_{R,t}, \\ E(\eta_{L,t}) &= 0, \text{Cov}(\eta_{L,t}, \eta_{L,t'}) = \begin{cases} \sigma_L^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} \\ E(\eta_{R,t}) &= 0, \text{Cov}(\eta_{R,t}, \eta_{R,t'}) = \begin{cases} \sigma_R^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \end{aligned} \quad (3.3)$$

The parameters L_t and R_t are referred to as the trend and the slope parameter respectively. The seasonal component is modelled with the trigonometric form

$$S_t = \sum_{l=1}^6 S_{l,t}, \quad (3.4)$$

where

$$S_{l,t} = S_{l,t-1} \cos(h_l) + S_{l,t-1}^* \sin(h_l) + \omega_{l,t}$$

$$S_{l,t}^* = S_{l,t-1}^* \cos(h_l) - S_{l,t-1} \sin(h_l) + \omega_{l,t}^*, \quad l = 1, \dots, 6,$$

$$h_l = \frac{\pi l}{6}, \quad l = 1, \dots, 6,$$

$$E(\omega_{l,t}) = E(\omega_{l,t}^*) = 0,$$

$$\begin{aligned} \text{Cov}(\omega_{l,t}, \omega_{l',t'}) &= \text{Cov}(\omega_{l,t}^*, \omega_{l',t'}^*) \\ &= \begin{cases} \sigma_\omega^2 & \text{if } l = l' \text{ and } t = t' \\ 0 & \text{if } l \neq l' \text{ or } t \neq t' \end{cases}, \end{aligned}$$

$$\text{Cov}(\omega_{l,t}, \omega_{l,t}^*) = 0 \text{ for all } l \text{ and } t. \quad (3.5)$$

The irregular component ε_t contains the unexplained variation and is modelled as a white noise process:

$$E(\varepsilon_t) = 0, \quad \text{Cov}(\varepsilon_t, \varepsilon_{t'}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \quad (3.6)$$

3.2 Time series model for rotation group bias

The systematic differences between the trend and the seasonal components of the subsequent waves are modelled in (3.1) with λ_t and γ_t . Additional restrictions for the elements of both vectors are required to identify model (3.1). Here it is assumed that an unbiased estimate for θ_t is obtained with the first wave, which is observed by CAPI, i.e., Y_t^1 . This implies that the first component of λ_t and γ_t equals zero. Now λ_t measures the systematic differences in the trend of the second, third, fourth and fifth wave with respect to the first wave. The components of λ_t are defined as:

$$\lambda_t^0 = 0, \quad \lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,j,t}, \quad j = 3, 6, 9, 12, \quad (3.7)$$

$$E(\eta_{\lambda,j,t}) = 0,$$

$$\text{Cov}(\eta_{\lambda,j,t}, \eta_{\lambda,j',t'}) = \begin{cases} \sigma_\lambda^2 & \text{if } t = t' \text{ and } j = j' \\ 0 & \text{if } t \neq t' \text{ or } j \neq j'. \end{cases}$$

Furthermore γ_t measures the systematic differences in the seasonal components with respect to the first wave. This implies that $\gamma_t^0 = 0$. The other components of γ_t are

defined as trigonometric functions, which are of the form of (3.5). The variance of the disturbances of the seasonal components are assumed to be equal for all waves and is denoted by σ_γ^2 .

To borrow information across the panel waves, the RGB for the trend as well as the RGB for the seasonal components are modelled as time invariant components, i.e., $\sigma_\lambda^2 = \sigma_\gamma^2 = 0$. As a kind of model diagnostic, the model initially allows for time dependent components. The maximum likelihood estimates for σ_λ^2 and σ_γ^2 are close to zero in this application. If this is not the case, it might be possible to allow for separate time independent RGB components for different time intervals.

3.3 Time series model for the survey errors

Finally a time series model for the survey errors in (3.1) is developed, which uses the direct estimates for the variance and AC's for the survey errors of the different panels as prior information. From (3.1) it follows that the survey errors for the first wave are defined as $e_t^1 = Y_t^1 - \theta_t$. For the second, third, fourth and fifth wave, they are defined as $e_t^{t-j} = Y_t^{t-j} - \theta_t - \lambda_t^j - \gamma_t^j$, for $j = 3, 6, 9, 12$.

Direct estimates for the variances of the survey errors for the separate panels are obtained with (2.2). These estimates are smoothed by modelling the variance estimates for the separate panels with a linear regression model $\text{Var}(Y_t^{t-j}) = b_0^j + b_1^j(Y_t^{t-j}/n_t^{t-j}) + \text{error}$, where n_t^{t-j} denotes the sample size at time t of the sample that entered the panel at $t-j$.

The rotating panel design implies sample overlap with panels observed in the past. The sample of the first wave enters the panel for the first time at time t , so there is no sample overlap with panels observed in the past. Consequently, the survey errors of the first wave, e_t^1 , are not correlated with survey errors in the past. The survey error of the second wave, i.e., e_t^{t-3} , is correlated with the survey error of the first wave that entered the panel three months earlier, i.e., e_{t-3}^{t-3} . In a similar way, the survey error of the third wave, i.e., e_t^{t-6} , is correlated with e_{t-3}^{t-6} and e_{t-6}^{t-6} . The survey error of the fourth wave, i.e., e_t^{t-9} , is correlated with e_{t-3}^{t-9} , e_{t-6}^{t-9} and e_{t-9}^{t-9} . Finally, the survey error of the fifth wave, i.e., e_t^{t-12} , is correlated with e_{t-3}^{t-12} , e_{t-6}^{t-12} , e_{t-9}^{t-12} and e_{t-12}^{t-12} .

The AC's between the survey errors of the subsequent waves are estimated using the approach proposed by Pfeiffermann *et al.* (1998). Since the real survey errors cannot be observed directly, this approach starts with calculating the autocovariances for the pseudo survey errors, which are defined as $(Y_t^{t-j} - \bar{Y}_t)$, where \bar{Y}_t denotes the average of the five panel estimates Y_t^{t-j} at time t . The autocovariances of the pseudo survey errors for a separate wave are influenced by the autocovariances of the real survey errors of the other waves, since the pseudo survey

errors are defined as the deviation of a panel estimate with the average of all panel estimates obtained at time t . Equation (4) of Pfeiffermann *et al.* (1998) specifies the relation between the autocovariances of the pseudo survey errors and the real survey errors. From this equation, it follows that the autocovariances of the real survey errors can be derived from the autocovariances of the pseudo survey errors by $\phi_k = F^{-1}C_k$, with C_k a vector containing the five autocovariances of the pseudo survey errors at lag k , ϕ_k a vector containing the five autocovariances of the survey errors at lag k , and F a $M \times M$ dimensional matrix where the diagonal elements equal $(M - 1/M)^2$ and the off-diagonal elements $(1/M)^2$. Here M denotes the number of waves of the panel design ($M = 5$ in this application). The AC's and the partial autocorrelations (PAC) of the survey errors of the subsequent waves are given in Table 3.1.

Table 3.1
Correlations and partial autocorrelations for the survey errors of the separate panels

wave		lag			
		1	2	3	4
1	AC	-0.029	0.264	0.022	0.230
	PAC	-0.029	0.263	0.038	0.175
2	AC	<u>0.291</u>	0.135	0.035	-0.250
	PAC	<u>0.291</u>	0.054	-0.020	-0.287
3	AC	<u>0.240</u>	<u>0.120</u>	0.087	0.219
	PAC	<u>0.240</u>	<u>0.066</u>	0.047	0.194
4	AC	<u>0.442</u>	<u>0.253</u>	<u>0.122</u>	0.156
	PAC	<u>0.442</u>	<u>0.072</u>	<u>-0.016</u>	0.115
5	AC	<u>0.249</u>	<u>0.298</u>	<u>-0.183</u>	<u>0.127</u>
	PAC	<u>0.249</u>	<u>0.252</u>	<u>-0.344</u>	<u>0.218</u>
Mean*	AC	0.306	0.224	-0.030	0.127
	PAC	0.306	0.144	-0.150	0.162

Underlined AC's and PAC's refer to waves with sample overlap

*) Means are based on the waves with sample overlap.

The standard errors of the estimated AC's equal $1/\sqrt{T}$, where T denotes the number of observations. This implies that correlations with an absolute value larger than 0.21 are significantly different from zero at a 5% significance level. The lags in Table 3.1 refer to three months periods, so lag one equals a time lag of three months, lag two a time lag of six months, *etc.*

The AC's in Table 3.1, which are based on overlapping samples, are underlined. The AC's for the overlapping samples are positive as might be expected. An exception is the AC at lag three for the fifth wave, which has a negative value. This correlation, however, is not significantly different from zero. The AC's for lag one of the overlapping samples are all significantly different from zero. For lag two, the AC's of the overlapping samples are significantly different from zero for the fourth and the fifth wave, but not

for the third wave. The AC's that are based on non-overlapping samples are sometimes unexpectedly large, *e.g.*, lag two and four of the first wave and lag four of the third wave. The AC for lag four of the second wave, on the other hand, has a surprisingly large negative value.

Pfeiffermann *et al.* (1998) also report large positive AC's for lags with non overlapping samples. In their case this can be explained since samples are replaced in small geographical regions. In the Dutch LFS sample replacement takes place at the national level. There is no good explanation why the AC's for the non overlapping samples are sometimes small and sometimes take significant positive as well as negative values. To obtain more stable estimates, the AC's are averaged over the waves which are based on overlapping samples. Thus the mean AC for lag one is the average of the AC for the second, third, fourth and fifth wave, *etc.* The values are reported in the last two rows of Table 3.1. The standard errors of the PAC's of order $p + 1$ and higher for an $AR(p)$ equal $1/\sqrt{T}$, Box and Jenkins (1970). This implies that the PAC's are not significantly different from zero for lags two and higher if an $AR(1)$ model with a correlation coefficient of 0.306 is assumed to capture the AC of the survey errors for the second, third, fourth and fifth wave.

The direct estimates for the variance and covariance structure of the survey errors are combined in the time series model using the following general form of the survey error model $e_t^{t-j} = k_t^{t-j} \tilde{e}_t^{t-j}$ where $k_t^{t-j} = \sqrt{\text{Var}(Y_t^{t-j})}$, see Binder and Dick (1990). This allows for non homogeneous variance in the survey errors, that arise *e.g.*, due to the gradually decreasing sample size over the last decade.

Since the first wave is uncorrelated with survey errors obtained in the past, it is assumed that \tilde{e}_t^t is white noise with $E(\tilde{e}_t^t) = 0$ and $\text{Var}(\tilde{e}_t^t) = 1$. As a result, the variance of the survey error equals $\text{Var}(e_t^t) = (k_t^t)^2$, which is equal to the direct estimate of the variance of the GREG estimate for the first wave. For the second, third, fourth and fifth wave, it is assumed that $\tilde{e}_t^{t-j} = \rho \tilde{e}_{t-3}^{t-j} + v_t^{t-j}$, with $\rho = 0.306$, and

$$E(v_t^{t-j}) = 0, \text{Cov}(v_t^{t-j}, v_{t'}^{t'-j}) = \begin{cases} \sigma_v^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases}$$

Since \tilde{e}_t^{t-j} is an $AR(1)$ process, $\text{Var}(\tilde{e}_t^{t-j}) = \sigma_v^2 / (1 - \rho^2)$. To enforce that $\text{Var}(e_t^{t-j})$ equals the direct estimate for the variance of the GREG estimate, it follows that $\sigma_v^2 = (1 - \rho^2)$.

3.4 Final time series model for the monthly unemployment rate

The time series model for the vector with GREG estimates Y_t is obtained by inserting the different components developed in Sections 3.1 through 3.3 into (3.1). This model

uses the five monthly GREG estimates as input data to obtain model-based estimates for the monthly unemployment rate. The component for the population parameter θ , in (3.2), developed in Section 3.1, takes advantages of sample information observed in the past to improve the precision of the estimated monthly unemployment rate. The components for the RGB, developed in Section 3.2, account for the systematic differences between the five monthly GREG estimates to avoid that the estimated monthly unemployment rate is incurred with this bias. The component for the survey errors, developed in Section 3.3, accounts for the AC between the five GREG estimates that are based on the same sample, observed with quarterly intervals. Although this approach is model-based, it accounts for the complexity of the survey design of the LFS, since the GREG estimates are used as input data.

4. State space representation

The time series model for the five monthly GREG estimates developed in Section 3 can be expressed in the state space representation, see Harvey (1989) or Durbin and Koopman (2001). A state space model consists of a measurement equation and a transition equation. The measurement equation, which is sometimes also called the signal equation, specifies how the observations depend on a linear combination of the state vector that contains the unobserved state variables for the trend, seasonal, RGB and the survey errors. The transition equation, which is sometimes also referred to as the system equation, specifies how the state vector evolves in time. The state space representation of the model developed in Section 3 is given by Van den Brakel and Krieg (2009).

Under the assumption of normally distributed error terms, the Kalman filter can be applied to obtain optimal estimates for the state vector. Estimates for state variables for period t based on the information available up to and including period t are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the completely observed time series. So the smoothed estimate for the state vector for period t also accounts for the information made available after time period t . In this paper, the Kalman filter estimates for the state variables are smoothed with the fixed interval smoother. See Harvey (1989), and Durbin and Koopman (2001) for technical details.

The analysis is conducted with software developed in Ox in combination with the subroutines of SsfPack 3.0, see Doornik (1998) and Koopman, Shephard and Doornik (2008). All state variables are non-stationary with the

exception of the survey errors. The non-stationary variables are initialised with a diffuse prior, *i.e.*, the expectation of the initial states are equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. The survey errors are stationary and therefore initialised with a proper prior. The initial values for the survey errors are equal to zero and the covariance matrix is available from the model developed for the survey errors in Section 3.3. In Ssfpack 3.0 an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997).

5. Results

5.1 Preliminary analyses

With the GREG estimator monthly estimates for the unemployment rate are obtained for each wave as described in Section 2.4. In Figure 5.1 the unemployment rate based on the CAPI wave is compared with the average of the four CATI waves. The graph shows that the unemployment rate observed with the first wave is systematically higher than for the other four waves.

The five time series obtained with the different waves are modelled with the time series model proposed in Sections 3 and 4. Preliminary analyses indicate that the estimates for the RGB of the seasonal effects in the second wave are not significantly different from zero and the RGB for the seasonal effects of the third, fourth and fifth wave are not significantly different from each other. Therefore the model is simplified to one with a single RGB seasonal effect. See Van den Brakel and Krieg (2009) for the state space representation.

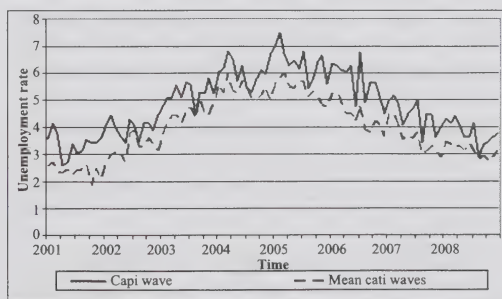


Figure 5.1 RGB monthly unemployment rate based on GREG estimates

5.2 Estimation results for the time series model

Maximum likelihood estimates for the hyperparameters, *i.e.*, the variance components of the stochastic processes for the state variables are obtained using a

numerical optimization procedure (BFGS algorithm, Doornik 1998). To avoid negative variance estimates, the log-transformed variances are estimated. The maximum likelihood estimates for the log-transformed variance of the level of the trend (σ_L^2), the seasonal component (σ_ω^2), the RGB of the trend (σ_λ^2) and the RGB of the seasonals tend (σ_χ^2) tend to large negative values with extremely large standard errors. These variance components are therefore put to zero in the final model. The estimation results for the remaining hyperparameters are presented in Table 5.1.

Table 5.1
Maximum likelihood estimates hyperparameters

Hyperparameter	Ln-transformed variance comp.		Variance components		
	Estimate	St. error	Estimate	95% conf. interval	
				Lower b.	Upper b.
Slope (σ_L^2)	-17.226	0.549	0.182E-3	0.106E-3	0.311E-3
Irregular comp. (σ_ϵ^2)	-13.480	0.482	1.183E-3	0.737E-3	1.897E-3

The smoothed Kalman filter estimates for the unemployment rate θ_t are given in Figure 5.2. These are the estimates for the monthly unemployment rate, based on the smooth trend model and a seasonal component, corrected for the RGB between the five GREG estimates. The local linear trend model simplified to a smooth trend model since $\sigma_L^2 = 0$. The trend component is time dependent since the maximum likelihood estimate of the hyperparameter for the slope is positive (see Table 5.1). The seasonal component is also time independent, since $\sigma_\omega^2 = 0$. Therefore the estimated seasonal effects obtained with the trigonometric form are exactly the same as the results obtained with the well known dummy variable seasonal model. The smoothed Kalman filter estimates for the trend and the seasonal component are plotted in Figures 5.3 and 5.4 respectively.

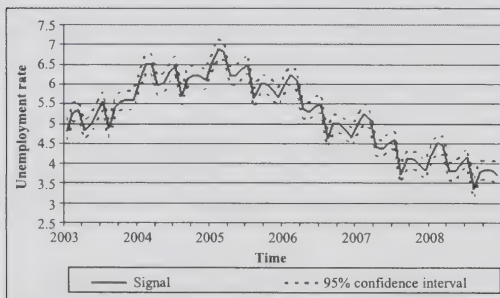


Figure 5.2 Smoothed Kalman filter estimates for the monthly unemployment rate

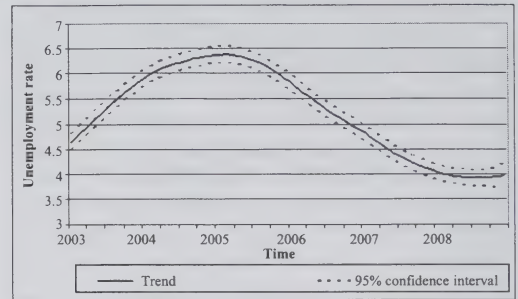


Figure 5.3 Smoothed Kalman filter estimates for the trend of the monthly unemployment rate

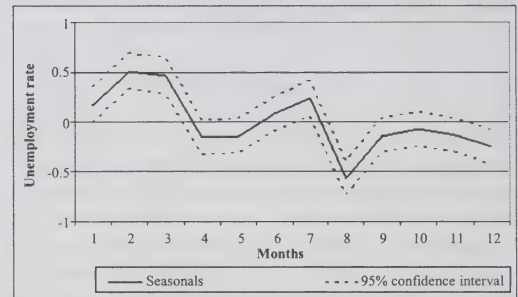


Figure 5.4 Smoothed Kalman filter estimates for the seasonal effect of the monthly unemployment rate

The Kalman filter estimates for the RGB of the trend are time independent. The smoothed Kalman filter estimates for the RGB are given in Table 5.2. The model beautifully detects a slightly increasing bias in the trend of the subsequent waves. The estimates for the RGB of the four CATI waves are significantly different from zero.

Table 5.2
Smoothed Kalman filter estimates RGB trend

Wave	RGB	St. error
2	-0.75	0.04
3	-0.86	0.04
4	-0.96	0.05
5	-1.10	0.05

An interesting empirical result of this application is the finding of the seasonality in the RGB. The Kalman filter estimates for the RGB of the seasonal effects are also time independent. Therefore, a sequence of likelihood ratio tests is conducted to reach the finally selected model and to test whether the seasonality effects in the RGB of this model are jointly significantly different from zero. Consider the following nested models:

- M1: separate and fixed RGB in the seasonality for wave two, three, four and five
- M2: equal to M1 where the RGB in the seasonality of wave two is equal to zero
- M3: equal to M2 with equal RGB in the seasonality of wave three, four and five
- M4: RGB in the seasonality of wave two, three, four and five is equal zero

The results of the likelihood ratio tests of this sequence of models are specified in Table 5.3.

Table 5.3
Likelihood-ratio tests for RGB in seasonality

Model	Log likelihood	Null hypothesis	Likh. ratio stat.	D.f.	p-value
M1	1,592.9				
M2	1,585.5	M2 = M1	14.7	11	0.19568
M3	1,573.7	M3 = M2	23.7	22	0.36422
M4	1,559.9	M4 = M3	27.6	11	0.00373

Testing the hypothesis that M2 equals M1 shows that the seasonality of the second wave is not significantly different from the first wave. Testing the hypothesis that M3 equals M2 shows that the RGB in the seasonality of the third, fourth and fifth wave are not significantly different. Testing the hypothesis that M4 equals M3 shows that the RGB of seasonal effects in last three waves are jointly significantly different from zero.

The smoothed Kalman filter estimates for the RGB of the seasonal effects for wave three, four and five are given in Figure 5.5. The smoothed Kalman filter estimates of the seasonal effects are compared with the smoothed estimates for the RGB of the seasonal effects in Figure 5.6.

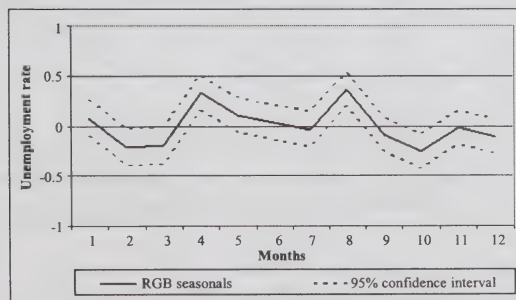


Figure 5.5 Smoothed Kalman filter estimates for the RGB of the seasonal effects in the third, fourth and fifth wave

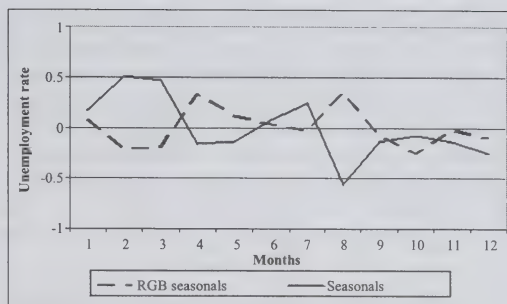


Figure 5.6 Comparison of smoothed Kalman filter estimates for the RGB of the seasonal effects in the third, fourth and fifth wave and the seasonal effects in 2008

It follows from Figure 5.5 that the seasonal effects in February, March, April, August and October in the third, fourth and fifth wave are significantly different from the first and the second wave. Figure 5.6 shows that the RGB in the seasonal effects largely nullifies the seasonal effects in these months. The seasonal effects in the last three waves are, apparently, less pronounced than in the first two waves. The different factors that contribute to the RGB in both the trend and the seasonal patterns are summarised in Section 2.2.

5.3 Comparison with GREG estimates

In this section the monthly GREG estimates for the unemployment rate and their standard errors are compared with the filtered model estimates. The filtered estimates are used since they are based on the complete set of information that would be available in the regular production process to produce a model-based estimate for the monthly unemployment rate for month t .

The GREG estimates based on the CAPI wave for the monthly unemployment rates are compared with the filtered model estimates in Figure 5.7. Some of the peaks and dips in the series of the GREG estimates are partially considered as survey errors under the structural time series model and flattened out in the filtered estimates for the series. Some of these peaks and dips are preserved since they are considered as seasonal effects under the time series model. It also follows that the filtered estimates are corrected for the RGB since the filtered series is at the same level as the series of the GREG estimates based on the CAPI wave. This is enforced with the assumption that the model parameters for the RGB for the first wave are zero (Section 3.2). This implies that the CATI waves are benchmarked to the outcomes of the first wave.

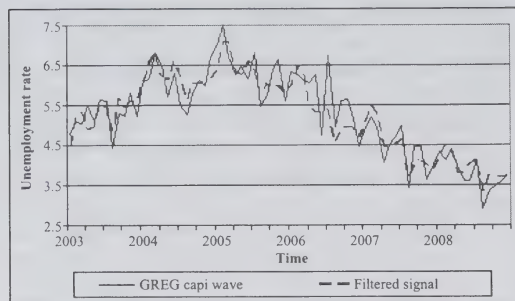


Figure 5.7 Filtered estimates and GREG estimates based on the CAPI wave for the monthly unemployment rate

The procedure applied in the regular estimation procedure of the LFS, to combine the CATI and the CAPI waves, is also used to estimate monthly unemployment figures. The GREG estimates for the monthly unemployment rates based on the five waves, using formula (2.4), are compared with the filtered estimates in Figure 5.8. Both estimates for the monthly unemployment rate follow the same level, since they are both benchmarked to the outcomes of the first wave. The GREG estimator is benchmarked in a rather rigid way using ratio (2.3), which is assumed to be constant in advance over a period of three years. The filtered estimates are benchmarked in a more subtle way through the explicit modelling of the trend and the seasonality in the RGB. The seasonality in the RGB indicates that the assumption of a constant RGB is not tenable. The monthly GREG estimates based on all waves are also compared with the GREG estimates based on the CAPI wave in Figure 5.9.

The ratio correction applied in formula (2.4) to the GREG estimates based on all waves removes the RGB in the trend, but does not correct for the RGB in the seasonal patterns. This follows from Figure 5.8 and 5.9. The series of the GREG estimates based on all waves follows the same level as the GREG estimates based on the CAPI wave (Figure 5.9). There are, however, subtle differences between the filtered estimates and the GREG estimates based on all waves (Figures 5.8). They partially arise because some of the dips and peaks in the GREG estimates are considered as survey errors by the time series model but they are also the result of systematic differences in the seasonal patterns between the subsequent waves. For example, the model estimates in February and March are larger in 2003, 2005 and 2006, and smaller in August in 2004, 2005 and 2006.

The standard errors of the monthly GREG estimates based on all waves, the CAPI wave and the filtered estimates are compared with each other in Figure 5.10. The standard errors for the GREG estimates are computed as

described in Section 2.4. Standard errors of the filtered estimates are obtained by the standard recursion formulas of the Kalman filter, see Harvey (1989) or Durbin and Koopman (2001). The Kalman filter recursion assumes that the fitted state space model is the truth. As a result the standard errors for the filtered estimates do not reflect the additional variation induced by the use of likelihood estimates for the variance components in the state space model and are therefore too optimistic.

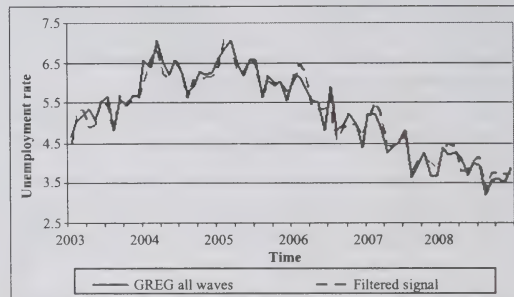


Figure 5.8 Filtered estimates and GREG estimates based on all waves for the monthly unemployment rate

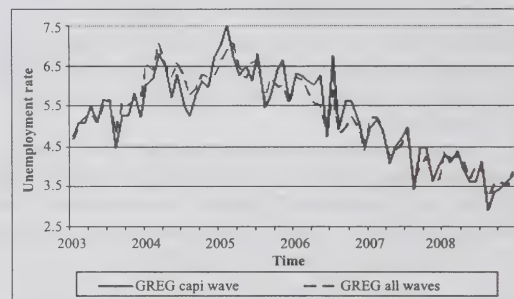


Figure 5.9 GREG estimates based on the CAPI wave and based on all waves for the monthly unemployment rate

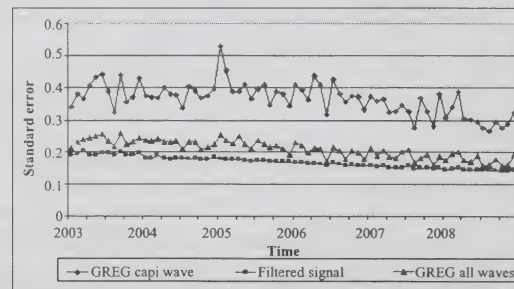


Figure 5.10 Standard errors of the GREG and filtered estimates for the monthly unemployment rate

As expected, the standard errors of the GREG estimates based on all waves are smaller than the standard errors of the GREG estimates based on the CAPI wave, since they are based on more data. The standard errors of the filtered estimates are smaller than the GREG estimates based on all waves, since the time series model uses additional sample information from preceding periods. The standard errors of the filtered estimates are slightly but continuously decreasing during the period 2003 to 2008.

The size and complexity of the applied time series model, is large compared to the length of the series available to fit the model. The final model that is applied to a five dimensional series which is monthly observed during a period of eight years contains 41 state variables. Therefore it is worthwhile to consider more parsimonious models, which might reduce the standard errors of the filtered estimates. Furthermore, the GREG estimate contains a bias since the RGB contains a seasonal effect, which is not reflected by its standard error. Therefore, the efficiency obtained by borrowing sample information from the past by relying on a time series model is illustrated more clearly if the standard error of the GREG estimates using all waves is compared with the standard error obtained with a time series model that accounts for the RGB in the trend only. Therefore a time series model without a component for the RGB in the seasonal pattern is applied to the data to illustrate the variance reduction by borrowing strength over time. The filtered estimates for the monthly unemployment rates based on a model with and without a component for the RGB in the seasonal pattern are compared in Figure 5.11.

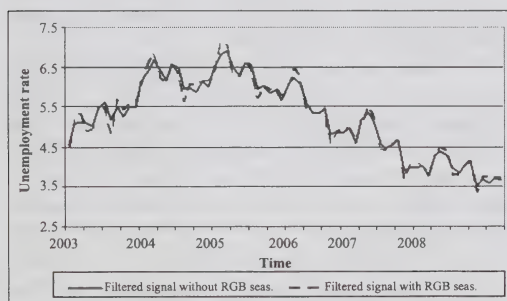


Figure 5.11 Filtered estimates of the monthly unemployment rate for two different time series models

The model without a component for the RGB of the seasonal effects assumes a seasonal effect for the population parameter θ_i that is based on an average of the seasonal effects of the five waves. The absolute values of the seasonal effects in February, March, and August are smaller under the simplified model, resulting in a lower estimate for

the monthly unemployment rate in February and March and a larger estimate in August. This results in a more pronounced seasonal pattern in the filtered series obtained with the complete model.

The standard errors of the filtered estimates obtained with the two time series models and the standard errors of the GREG estimates using all waves are compared in Figure 5.12. The standard error of the filtered estimates of the simplified time series model is substantially smaller than the standard error of the GREG estimates using all waves. The simplification of the time series model by ignoring the RGB for the seasonal effects, results in a further reduction of the standard error at the cost of an increased bias in the seasonal effects. Under the model assumption that the estimates based on the first wave are unbiased, the time series model that accounts for the RGB in the seasonal effects is preferred, since it removes the bias in the seasonal pattern.

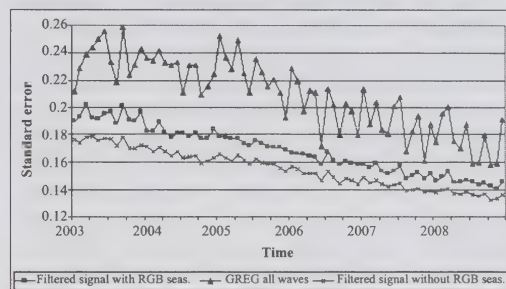


Figure 5.12 Standard errors of the GREG estimates based on all waves and filtered estimates for two different time series models for the monthly unemployment rate

Discussion

In this paper a multivariate structural time series model is applied to the monthly data of the LFS that accounts for the rotating panel design of this survey. This approach is initially proposed by Pfeffermann (1991) and extended in this paper with a component that models systematic differences in the seasonal effects between the subsequent waves. Compared with the GREG estimator, which is currently applied in the regular LFS, the time series model results in a substantial increase of the accuracy of the estimates of the unemployment rate. Firstly, the model explicitly estimates the RGB in the trend and the seasonal patterns between the first CAPI wave and the four subsequent CATI waves. Secondly, the time series model borrows strength from data observed in preceding periods via the assumed model for the population parameter and the AC between the survey errors of the different panels.

The RGB induced by the rotating panel design is substantial. The bias in the trend results in an under-estimation of the unemployment rate in the subsequent waves and its magnitude slightly decreases from -0.8 percent points in the second wave to -1.1 percent points in the fifth wave. The seasonal patterns of the first two waves and the last three waves are also significantly different, since the seasonal pattern in the last three waves is less pronounced.

A parsimonious time series model that accounts for the RGB in the trend but not for the RGB in the seasonal pattern, results in a further reduction of the standard error of the filtered estimates. This, however, results in a biased seasonal pattern in the monthly estimates of the unemployment rates. Since the standard errors of the filtered estimates obtained under this parsimonious model do not reflect this bias, a time series model that accounts for both the RGB in the trend and the seasonal pattern is preferred.

The time series model is identified by adopting a restriction for the RGB parameters which assumes that the first wave is observed without bias. This implies that the estimates based on the first wave are used to benchmark the subsequent waves. If this restriction is used, then an all out effort in each part of the statistical process is required to reduce possible bias in the first wave, *e.g.*, by using the most appropriate data collection mode, reducing nonresponse, optimizing the weighting scheme, *etc.* Based on external information about the bias in the different waves, the restrictions for the RGB components might be adjusted.

The time series approach explored in this paper is appropriate to produce model-based estimates for monthly unemployment figures. Statistics Netherlands, however, is generally rather reserved in the application of model-based estimation procedures for the production of official statistics. Model misspecification might result in severely biased estimates. This bias is not reflected in the standard errors of the Kalman filter estimates. Extensive model selection and evaluation is therefore required for each separate target variable. This hampers a straightforward application of such estimation techniques, since there is generally limited time available for the analysis phase of the regular production process of official releases.

There is, on the other hand, a case for having official series that are based on model-based procedures with appropriate methodology and quality descriptions for situations where direct estimators do not result in sufficiently reliable estimates. The RGB observed under the rotating panel design of the LFS clearly illustrates the existence of non-sampling errors such as measurement errors and panel attrition. Therefore the traditional concepts that observations obtained from sampling units are true fixed values observed without error and that the respondents

can be considered as a representative probability sample from the target population, generally assumed in design-based sampling theory, are not tenable under such designs. The application of direct estimators in the case of measurement errors and selective panel attrition will result in severely biased estimates. In the regular estimation procedure a ratio correction is applied to the GREG estimates, which is based on the implicit model assumption that the bias is constant over a period of three years. The time series model applied in this paper can be used to produce estimates that are corrected for the bias introduced by these non-sampling errors in a more advanced way.

This estimation procedure is also applicable in situations where small sample sizes result in unacceptable large standard errors. Small sample sizes arise if official statistics are required for small domains or for short data collection periods like the monthly unemployment figures in the LFS. Most surveys conducted by national statistical institutes operate continuously in time and are based on cross-sectional or rotating panel designs. Consequently, estimation procedures based on time series models that use sample information observed in preceding periods are particularly interesting.

The time series model yields estimates for the trend and seasonal components of the population parameter. Seasonally adjusted parameter estimates and their estimation errors are therefore obtained as a by-product of this estimation procedure. Another major advantage is that this approach accounts for the AC in the survey errors due to the rotating panel design. Pfeiffermann *et al.* (1998) show that ignoring these AC, for example with the Henderson filters in X-12-ARIMA (Findley, Monsell, Bell, Otto and Chen 1998), results in spurious trend estimates.

The model can be improved in several ways. Information about registered unemployment and related variables, available in the register of the Office for Employment and Income, can be used as auxiliary variables in the models. If longer series become available, an additional cyclic component might be required to capture economic fluctuations. Another possible improvement is detection and modelling of outliers. Furthermore the model needs to be extended to estimate monthly unemployment rates for different domains using sample information collected in the past as well as cross-sectional data from other small areas, using the approach proposed by Pfeiffermann and Burck (1990) and Pfeiffermann and Tiller (2006).

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank Professor

D. Pfeffermann and Professor S.J. Koopman for their valuable advice during this project as well as the Associate Editor and the referees for giving constructive comments on earlier drafts of this paper.

References

- Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- Bell, W.R., and Hillmer, S.C. (1990). The time series approach to estimation of periodic surveys. *Survey Methodology*, 16, 195-215.
- Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- Binder, D.A., and Dick, J.P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- Binder, D.A., and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 239-253.
- Box, G.E.P., and Jenkins, G.W.M. (1970). *Time series analysis - forecasting and control*. San Francisco: Holden-Day.
- Cantwell, P.J. (1990). Variance formulae for composite estimators in rotating designs. *Survey Methodology*, 16, 153-163.
- Doomik, J.A. (1998). *Object-oriented matrix programming using Ox 2.0*. London: Timberlake Consultants Press.
- Durbin, J., and Koopman, S.J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.
- Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55, 182-199.
- Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. and Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-176 (with Discussion).
- Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian labour force survey. *Survey Methodology*, 27, 45-51.
- Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27, 65-74.
- Gurney, M., and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics*, American Statistical Association, 242-257.
- Hansen, M.H., Hurwitz, W.N. and Meadow, W.G. (1953). *Sample survey methods and theory*, 2. New York: John Wiley & Sons, Inc.
- Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.
- Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.
- Holbrook, A.L., Green, M.C. and Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly*, 67, 79-125.
- Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Section on Social Statistics*, American Statistical Association, 300-303.
- Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.
- Koopman, S.J., Shephard, N. and Doornik, J.A. (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*. London: Timberlake Consultants Press.
- Kumar, S., and Lee, H. (1983). Evaluation of composite estimation for the Canadian labour force survey. *Survey Methodology*, 9, 178-201.
- Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- Nieuwenbroek, N., and Boonstra, H.J. (2002). *Bacula 4.0 reference manual*, BPA nr: 279-02-TMO, Statistics Netherlands, Heerlen.
- Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.
- Pfeffermann, D., and Rubin-Bleuer, S. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 149-163.
- Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, 339-348.
- Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.
- Rao, J.N.K., and Graham, J.E. (1964). Rotating designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.
- Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. Review paper, NCRM/008, National Centre for Research Methods, City University London.
- Särndal, C.-E., and Lundström, S. (2005). *Estimation in surveys with nonresponse*. New York: John Wiley & Sons, Inc.
- Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

- Särndal, C-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer Verlag.
- Scott, A.J., and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.
- Singh, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 120-129.
- Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian labour force survey with a rotating panel design. *Survey Methodology*, 27, 33-44.
- Tam, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.
- Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.
- van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch labour force survey. *Journal of the Royal Statistical Society, Series A*, 171, 581-613.
- van den Brakel, J.A., and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. Research paper, Statistics Netherlands, Heerlen (<http://www.cbs.nl/en-GB/menu/methoden/research/discussionpapers/archief/2009/default.htm?Languageswitch=on>).

Estimates for small area compositions subjected to informative missing data

Li-Chun Zhang¹

Abstract

Estimation of small area (or domain) compositions may suffer from informative missing data, if the probability of missing varies across the categories of interest as well as the small areas. We develop a double mixed modeling approach that combines a random effects mixed model for the underlying complete data with a random effects mixed model of the differential missing-data mechanism. The effect of sampling design can be incorporated through a quasi-likelihood sampling model. The associated conditional mean squared error of prediction is approximated in terms of a three-part decomposition, corresponding to a *naïve* prediction variance, a positive correction that accounts for the hypothetical parameter estimation uncertainty based on the latent complete data, and another positive correction for the extra variation due to the missing data. We illustrate our approach with an application to the estimation of Municipality household compositions based on the Norwegian register household data, which suffer from informative under-registration of the dwelling identity number.

Key Words: Conditional MSE of prediction; EMPQL algorithm; Generalized SPREE; Not missing-at-random; Two-way contingency table.

1. Introduction

Small area (or domain) population counts cross-classified by various social-economic characteristics are increasingly demanded for fund allocation, regional planning and social-economic research. Purcell and Kish (1980) outlined the so-called "Structure preserving estimation" (SPREE), which operates by modifying the small area estimates in a way so that they vary from one area to another in accordance with the variation that exists in another known auxiliary table of the same dimension. Typically the auxiliary table is obtained from a previous census, or some administrative register containing similar information. Zhang and Chambers (2004) developed a generalized SPREE (GSPREE) approach. Both fixed effects and random effects mixed models were introduced, and the restricted log-linear model underlying SPREE was shown to be a special case. This provides means for reducing the potential bias of the traditional SPREE estimates. We refer to Ghosh, Natarajan, Stroud and Carlin (1998) and Longford (1999) for alternative hierarchical and empirical Bayes approaches to this type of data.

In this paper we extend the GSPREE approach to situations subjected to missing data. This can be useful in sample surveys where nonresponse is unavoidable. We concentrate on *small area compositions* that can be arranged in a two-way table, where one of the two dimensions refers to the small areas and the other refers to the categories of interest. The cell counts summarize to a fixed area total that may or may not be known. For instance, each person between 16 and 74 years of age can be classified according to the labour force status "employed", "unemployed" and "not in the labour force". The sum of the three counts inside

a small area is the total number of persons between 16 and 74 years of age within this area.

In the context of small area composition we say that the missing-data mechanism is *informative* provided it varies across the categories of interest. As such it is also *not* missing-at-random (Rubin 1976). In addition, the overall rate of missing differs across the areas. Differential missingness as such leads to distortion of the underlying complete data, and bias if the estimation is carried out as if the observed data were complete. We propose a double mixed modeling approach that combines the random effects mixed model for the underlying complete data with a random effects mixed model of the missing-data mechanism. The double-smoothing approach is outlined in Section 2.

It should be noted that national statistical offices that conduct large scale surveys will have accounted for missing data by weighting adjustments or imputation. This, however, will have been done at levels that are significantly higher than the small areas, and will be for variables that do not necessarily correspond to those of interest for the small areas. When available, the adjusted totals can be incorporated into the GSPREE as marginal totals for iterative proportional fitting (IPF). But modeling of the differential probabilities of missing across the small areas will generally remain a matter of interest.

It should also be noticed that informative missing data as such makes it less straightforward to assess the potential bias of any estimation approach. SPREE may be biased on two accounts: (i) the underlying restricted log-linear assumptions are likely to be unrealistic, (ii) direct IPF may fail to account for the differential probabilities of missing

adequately. The proposed double mixed modeling approach deals with problem (i) by GSPREE modeling of the underlying complete data, and it deals with problem (ii) by introducing a more flexible missing-data model, as we shall discuss in Section 2.2. Nevertheless, bias is likely to persist to a certain extent. Since the estimation of model parameters and random effects is more complicated under the double mixed modeling approach, alternative estimation methods that are able to preserve the computational simplicity of SPREE, while making more adequate adjustment for informative missing data, are worth investigating in future.

When it comes to the assessment of estimation uncertainty, Booth and Hobert (1998) argued for the conditional mean squared error of prediction (CMSEP) given the observed data. We extend their approach and derive approximate CMSEP in the current multivariate incomplete-data situation. This results in a three-part decomposition of the CMSEP, corresponding to a *naive* prediction variance, a positive correction that accounts for the hypothetical estimation uncertainty of the parameters based on the latent complete data, and another positive correction for the extra variation due to the missing data. The details are given in Section 3.

Estimation procedures for the parameters, the CMSEP and the small area compositions are described in Section 4. In Section 5 we apply our approach to derive estimates of the Municipality household compositions based on the Norwegian household register, which suffers from informative under-registration of the dwelling identity number (DIN). A summary is given in Section 6.

2. Double mixed modeling

2.1 Random effects mixed model in the complete-data case

2.1.1 Models for finite population

The small area counts can be arranged in a two-way contingency table, denoted by $\mathbf{X} = \{X_{ak}\}$, where $a = 1, \dots, A$ indexes the small areas and $k = 1, \dots, K$ the categories of interest. The interest of estimation is the within-area proportions given by

$$\theta_{ak}^X = X_{ak} / X_{a\cdot} = X_{ak} / \sum_{j=1}^K X_{aj}$$

referred to as compositions since $\sum_k \theta_{ak}^X = 1$. Typically under the GSPREE approach we assume that the marginal totals $\{X_{a\cdot}\}$ and $\{X_{\cdot k}\}$, also known as the allocation structure, are either known or can be reliably estimated, in which case estimating $\{\theta_{ak}^X\}$ is equivalent to estimating $\{X_{ak}\}$. For simplicity we then make no distinction between counts and compositions in the exposition. Otherwise,

without the allocation structure, one can still use our approach to estimate $\{\theta_{ak}^X\}$ but not $\{X_{ak}\}$.

Assume that we have available an auxiliary table of the same dimension, denoted by $\mathbf{X}^0 = \{X_{ak}^0\}$, and the corresponding within-area proportions $\{\theta_{ak}^0\}$. To model $\theta_a^X = (\theta_{a1}^X, \dots, \theta_{aK}^X)^T$ we use the *multinomial standardized-log (mslog)* link function, given by

$$\mu_{ak}^X = \log \theta_{ak}^X - K^{-1} \sum_{j=1}^K \log \theta_{aj}^X \quad (1)$$

and similarly for μ_{ak}^0 and θ_{ak}^0 . Zhang and Chambers (2004) introduced the following generalized linear structural mixed model (GLSMM)

$$\mu_{ak}^X = \lambda_k + \beta \mu_{ak}^0 + v_{ak} \quad (2)$$

where

$$\sum_{k=1}^K \lambda_k = 0 \quad \text{and} \quad \sum_{k=1}^K v_{ak} = 0$$

and $\mathbf{v}_{a(1)} = (v_{a2}, \dots, v_{aK})^T$ assumes a multivariate normal distribution with covariance matrix $G = G(\delta)$, where δ contains the variance parameters. Notice that there is no area-specific term in (2) because $\sum_k \mu_{ak} = \sum_k \mu_{ak}^0 = 0$. The term “structural” refers to the fact that this is a model of the finite-population parameters $\{\theta_{ak}^X\}$ directly, although the emphasis is not common in the small area estimation literature. For instance, the well-known Fay-Herriot model (Fay and Herriot 1979) is “structural” in the same sense.

There is an important interpretation of the model (2) in terms of the log-linear interactions of $\{\theta_{ak}\}$ due to the choice of the link function (1), i.e.,

$$\mu_{ak}^X = \alpha_k + \alpha_{ak}^X \quad (3)$$

where by the standard theory of log-linear models (e.g., Agresti 2002), we have

$$\log X_{ak} = \log X_{a\cdot} + \log \theta_{ak}^X = \alpha_0^X + \alpha_a^X + \alpha_k^X + \alpha_{ak}^X$$

for $\alpha_0^X = (AK)^{-1} \sum_{a,k} \log X_{ak}$, and $\alpha_a^X = K^{-1} \sum_k \log X_{ak} - \alpha_0^X$, and $\alpha_k^X = A^{-1} \sum_a \log X_{ak} - \alpha_0^X$, and $\alpha_{ak}^X = \log X_{ak} - \alpha_a^X - \alpha_k^X - \alpha_0^X$, such that $\sum_a \alpha_a^X = \sum_k \alpha_k^X = \sum_a \alpha_{ak}^X = \sum_k \alpha_{ak}^X = 0$. We refer to (3) as the log-linear identity, and we refer to the log-linear parameters α_{ak}^X as the (first-order) interactions of the compositions θ_{ak}^X as well as the counts X_{ak} . Similar identity holds for μ_{ak}^0 . Zhang and Chambers (2004) showed that the GLSMM is equivalent to the following *proportional interactions mixed model (PIMM)*

$$\alpha_{ak}^X = \beta \alpha_{ak}^0 + v_{ak} + O_p(A^{-1/2}). \quad (4)$$

The parameters λ_k 's in (2) do not entail any model restriction beyond the PIMM, and they do not affect the

interactions. The parameter β is called the proportionality coefficient. Clearly, SPREE based directly on the association structure $\{X_{ak}^0\}$ amounts to setting $\beta \equiv 1$ and $v_{ak} \equiv 0$. We therefore refer to the model (2) as a GSPREE model, which contain both fixed and random effects extensions of the SPREE model.

2.1.2 Model for sample

To complete the model specification we assume sample classifications $\mathbf{x} = \{x_{ak}\}$. Let

$$\mathbf{t}_a = (t_{a1}, \dots, t_{aK})^T = (t_1(\mathbf{x}_a), \dots, t_K(\mathbf{x}_a))^T$$

be such that $E(t_{ak} | \mathbf{v}) = E(t_{ak} | \mathbf{X}) = \theta_{ak}^X$, where $\mathbf{v} = \{v_{ak}\}$. The expectation is typically with respect to the sampling design. However, it can also be taken under a suitable model of the sampling distribution, such as a multinomial model for \mathbf{x}_a provided simple random sampling within each area. We therefore make no distinction in the notation.

We assume that \mathbf{t}_a is independent of $\mathbf{t}_{a'}$ for $a \neq a'$, and put

$$V(t_{ak}) = v_1 \omega_k(\mathbf{X}_a) \quad \text{and} \quad \text{Cov}(t_{ak}, t_{aj}) = v_1 \omega_{kj}(\mathbf{X}_a) \quad (5)$$

where $\omega_k(\cdot)$ and $\omega_{kj}(\cdot)$ are specified variance and covariance functions, and v_1 is the dispersion parameter that may or may not be known. This is essentially the quasi-likelihood set-up for dependent data (McCullagh and Nelder 1989). The dependence on \mathbf{X}_a allows us to incorporate the sampling design effect, in which case the expectations in (5) may be evaluated with respect to the sampling distribution. This is an important reason why we do not directly assume that the distribution of \mathbf{t}_a belongs to the exponential family, as e.g., in the generalized linear mixed models (Breslow and Clayton 1993).

2.1.3 Parameter estimation

Zhang and Chambers (2004) outline an iterative weighted least square (IWLS) algorithm for the GLSMM (2), which is a variation of the PQL approach (Schall 1991; Breslow and Clayton 1993). Let $\mu_a = (\mu_{a1}^X, \dots, \mu_{aK}^X)^T$. The GLSMM (2) can formally be given by

$$\mu_a = g(\theta_a) = H_a \zeta + B \mathbf{v}_{a(1)}$$

where $g(\theta_a)$ is the mslog link function, and $\zeta = (\lambda_2, \dots, \lambda_K, \beta)^T$, and $\mathbf{v}_{a(1)} = (v_{a2}, \dots, v_{aK})^T$. The $K \times K$ design matrix H_a and $K \times (K-1)$ design matrix B are, respectively,

$$H_a = [B_{K \times K-1} \quad \mu_a^0] \quad \text{and} \quad B = \begin{pmatrix} -\mathbf{1}_{K-1}^T \\ I_{K-1 \times K-1} \end{pmatrix}$$

where $\mathbf{1}$ is a vector of 1 and I is an identity matrix. Define the working variables

$$\stackrel{\text{def.}}{\mathbf{z}_a} = \mu_a + \mathbf{e}_a = H_a \zeta + B \mathbf{v}_a + \mathbf{e}_a \quad \text{and} \quad \mathbf{e}_a = Q(\mathbf{t}_a - \theta_a^X) \quad (6)$$

where $Q = \partial \mu_a^X / \partial \theta_a^X$ is the Jacobian matrix of partial derivatives. Denote by R_a the conditional covariance matrix of \mathbf{t}_a given θ_a^X defined by (5). Under the PQL approach we assume that \mathbf{e}_a has an approximate multivariate normal distribution with covariance matrix $Q R_a Q^T$, and apply standard methods for linear mixed models (LMM) to the linearized data (6). Variants of the PQL approach differ in the estimation of the variance parameters δ . The details are omitted here.

2.1.4 On model hierarchy

The GLSMM (2) is specified at the finite population level. More generally, we may consider the finite population $\{X_{ak}\}$ to be randomly generated from an infinite super-population. Let θ_{ak} be the within-area probability that a unit of the super-population belongs to the cell (a, k) , where $\sum_k \theta_{ak} = 1$. Conditional on $X_a = \sum_k X_{ak}$, the within-area counts $(X_{a1}, \dots, X_{aK})^T$ follow the multinomial distribution with parameters $(\theta_{a1}, \dots, \theta_{aK})^T$. A *multinomial standardized-log mixed model (MSLMM)* of $\{\theta_{ak}\}$ is given by

$$\mu_{ak} = \lambda_k + \beta \mu_{ak}^0 + v_{ak} \quad (7)$$

where

$$\sum_{k=1}^K \lambda_k = 0 \quad \text{and} \quad \sum_{k=1}^K v_{ak} = 0$$

where μ_{ak} is given by θ_a through the mslog link function.

Unlike the GLSMM (2), the equation (7) defines a regression model. There are then three different hierarchy one may choose from in the sample survey situation:

1. Assume the GLSMM (2) for the finite population and the quasi-likelihood model (5) for the sample, yielding the GSPREE approach of Zhang and Chambers (2004).
2. Assume the MSLMM (7) for the super-population and model sample data \mathbf{t}_a based on θ_a directly, yielding a purely model-based two-level approach.
3. Assume the MSLMM (7) for the super-population, and assume that the finite population totals \mathbf{X}_a follow the multinomial distribution given θ_a , and assume the quasi-likelihood model (5) given \mathbf{X}_a , yielding a general three-level model.

Provided the finite population is large, it makes little difference in practice to adopt the GSPREE approach, in

which case one does not have to deal explicitly with one extra level of hierarchy. But the distinction between (2) and (7) becomes necessary if the areas are so small that the stochastic variation in \mathbf{X}_a is not negligible compared to the sampling variation in \mathbf{x}_a (or \mathbf{t}_a). In our application later, we have register data that would have given us the interested population counts $\{X_{ak}\}$ had they not suffered from missing data. And the small area level of aggregation is so detailed that the stochastic variation in \mathbf{X}_a can not be ignored. We therefore adapt the GSPREE approach by (a) adopting the MSLMM (7) instead of the GLSMM (2), and (b) modeling \mathbf{X}_a as a 'sample', albeit a very large one, from the super-population directly.

2.2 A random effects mixed model of missing data

Missing data add another level of stochastic variation on top of the underlying complete data. In the exposition below, we consider the sample counts $\{x_{ak}\}$ as the complete data, which is the most common situation in practice. Our application later in Section 5 can be viewed as a special case where $\mathbf{X} = \mathbf{x}$.

Denote by $\mathbf{y}_a = (y_{a1}, \dots, y_{aK})^T$ the observed cell counts, for $a = 1, \dots, A$. Suppose that, conditional on x_{ak} and a random effect b_a ,

$$E(y_{ak} | x_{ak}, b_a) = x_{ak} p_{ak} \quad (8)$$

$$V(y_{ak} | x_{ak}, b_a) = v_2 c_{ak} p_{ak} (1 - p_{ak})$$

where c_{ak} is a known constant, and v_2 is the dispersion parameter. We assume that y_{ak} is independent of y_{aj} for $k \neq j$, i.e., missing data are independent from one cell to another. Let the units in the complete sample cell (a, k) be indexed by $i = 1, \dots, n_{ak}$. Let $r_{i,ak} = 1$ if the i^{th} unit is observed, and $r_{i,ak} = 0$ if it is missing. The parameter p_{ak} is the assumed probability of $r_{i,ak} = 1$ inside cell (a, k) . To see this, let $x_{i,ak}$ be the contribution of the i^{th} unit to x_{ak} , i.e., $x_{ak} = \sum_{i=1}^{n_{ak}} x_{i,ak}$, such that $y_{ak} = \sum_{i=1}^{n_{ak}} r_{i,ak} x_{i,ak}$ and

$$E(y_{ak} | x_{1,ak}, \dots, x_{n_{ak},ak}, b_a) = \sum_{i=1}^{n_{ak}} x_{i,ak} E(r_{i,ak} | b_a) = \sum_{i=1}^{n_{ak}} x_{i,ak} P(r_{i,ak} = 1 | b_a) = x_{ak} p_{ak}.$$

Notice that p_{ak} does not depend on the value of $x_{i,ak}$, but only the position of the unit in the two-way table. We assume that p_{ak} depends on b_a through the logistic link function given by

$$\eta_{ak} = \log(p_{ak}/(1 - p_{ak})) = \xi_k + b_a \quad (9)$$

where

$$b_a \sim N(0, \sigma^2).$$

The fixed effects ξ_k 's allow the probability of missing to depend on the categories of interest, the area-level random effect b_a allows it to vary across the areas in addition.

Obviously, under the assumptions (8) and (9), the missing data cause bias in the estimates of the λ_k 's, if the observed table \mathbf{y} is treated as if it were complete. Moreover, it distorts the estimation of the first-order interactions $\{\alpha_{ak}^x\}$. We have,

$$\log p_{ak} = (\xi_k + b_a) - \gamma_{ak} \quad \text{where } \gamma_{ak} = \log(1 + \exp(\xi_k + b_a)).$$

The first-order interactions of $\{p_{ak}\}$ are then given by $\alpha_{ak}^p = -\tilde{\gamma}_{ak} = -(\gamma_{ak} - \bar{\gamma}_a - \bar{\gamma}_k + \bar{\gamma})$, for the row and column means $\bar{\gamma}_a$ and $\bar{\gamma}_k$ and the overall mean $\bar{\gamma}$. These are non-zero unless $\xi_k = \xi$. By (8) the interactions of the expected observed table are given by

$$\alpha_{ak}^{E(\mathbf{y}|\mathbf{x}, \mathbf{b})} = \alpha_{ak}^x + \alpha_{ak}^p = \alpha_{ak}^x - \tilde{\gamma}_{ak} \neq \alpha_{ak}^x$$

such that the estimates of $\{\alpha_{ak}^x\}$ will be biased if \mathbf{y} is treated as \mathbf{x} .

It is worth noting that, as far as the estimation of the interactions is concerned, it is in principle possible to treat the observed table \mathbf{y} as if it were the complete table \mathbf{x} under a particular missing-data model given by

$$\log p_{ak} = \xi'_k + b'_a. \quad (10)$$

This is because the first-order interactions of $\{p_{ak}\}$ are all zero under (10), in which case we have $\alpha_{ak}^{E(\mathbf{y}|\mathbf{x})} = \alpha_{ak}^x$. Disregarding the range restrictions, the assumption (10) defines an informative missing-data mechanism where the probability of missing varies across the categories of interest, while the area effect modifies all the within-area probabilities of missing by a factor $\exp(b'_a)$, such that $p_{ak} / \sum_{j=1}^K p_{aj} = \exp(\xi'_k) / \sum_j \exp(\xi'_j)$ remains constant. The model (9), however, is more flexible since it allows the random effects to affect the interactions. Both (9) and (10) will be examined in Section 5.

Finally, we notice that allowing for component-wise random effects in the model (9) may cause identification problems. For instance, assume simple random sampling from the finite population, in which case the interactions of the expected complete table are given by $\alpha_{ak}^{E(\mathbf{x}|\mathbf{X})} = \alpha_{ak}^x$. With component-wise b_{ak} in the model (9) we have $\log p_{ak} = \xi_k + b_{ak} + \gamma_{ak}$, where $\gamma_{ak} = \log(1 + \exp(\xi_k + b_{ak}))$. It follows from (4) and (8) that the interactions of the expected table $E(\mathbf{y}|\mathbf{x}, \mathbf{b})$ is given by $\beta \alpha_{ak}^0 + \gamma_{ak} + b_{ak} - \tilde{\gamma}_{ak}$. But there is no information in the observed data to distinguish between the two random effects v_{ak} and b_{ak} .

3. Conditional mean squared errors of prediction

We adopt the approach of Booth and Hobert (1998) and use the CMSEP as a measure of the uncertainty in prediction. Like them we consider the CMSEP on the linear-predictor scale. In vector form the μ_{ak} 's in (1) belong to the following class of linear functions

$$\mu_a = H_a \zeta + B_a \mathbf{v}_a \quad (11)$$

where μ_a is the area-specific vector of linear predictors, and ζ is the vector of fixed effects, and \mathbf{v}_a is the vector of area-specific random effects, and H_a and B_a are the corresponding design matrices. All the quantities have been specified in (6) for the GLSMM (2), where we actually have $B_a = B$. But we shall adopt the slightly more general formulation (11) in the following. Let $\hat{\zeta}$ and $\hat{\mathbf{v}}_a$ be, respectively, the estimates of ζ and \mathbf{v}_a based on observations subjected to missing data, denoted by \mathbf{y}_a for $a = 1, \dots, A$. The CMSEP of $\hat{\mu}_a = H_a \hat{\zeta} + B_a \hat{\mathbf{v}}_a$ is defined as

$$\text{CMSEP}_a = E\{(\hat{\mu}_a - \mu_a)(\hat{\mu}_a - \mu_a)^T | \mathbf{y}_a\}.$$

We introduce first a decomposition through the hypothetical best predictor (BP) based on \mathbf{x}_a , given by $\hat{\mu}_a = E(\mu_a | \mathbf{x}_a, \zeta, \delta) = H_a \zeta + B_a E(\mathbf{v}_a | \mathbf{x}_a, \zeta, \delta)$, when the parameters are known. We have

$$\begin{aligned} \text{CMSEP}_a &= E\{E((\hat{\mu}_a - \mu_a)(\hat{\mu}_a - \mu_a)^T | \mathbf{x}_a) | \mathbf{y}_a\} \\ &\quad + E\{(\hat{\mu}_a - \hat{\mu}_a)(\hat{\mu}_a - \hat{\mu}_a)^T | \mathbf{y}_a\} \\ &= E\{B_a \text{Cov}(\mathbf{v}_a, \mathbf{v}_a | \mathbf{x}_a) B_a^T | \mathbf{y}_a\} \\ &\quad + E\{(\hat{\mu}_a - \hat{\mu}_a)(\hat{\mu}_a - \hat{\mu}_a)^T | \mathbf{y}_a\} \end{aligned}$$

because $\hat{\mu}_a - \mu_a$ and $\hat{\mu}_a - \hat{\mu}_a$ are conditionally independent of each other given \mathbf{x}_a : $\hat{\mu}_a - \mu_a$ depends on the random effects \mathbf{v}_a , whereas $\hat{\mu}_a - \hat{\mu}_a$ depends on random variations in the other areas. Next, for the second term on the right-hand side, we introduce a decomposition through the hypothetical estimated best predictor (EBP) based on the complete data \mathbf{x} , denoted by $\tilde{\mu}_a = H_a \tilde{\zeta} + B_a \tilde{\mathbf{v}}_a$, where $(\tilde{\zeta}, \tilde{\delta})$ are the parameter estimates based on \mathbf{x} , and $\tilde{\mathbf{v}}_a = E(\mathbf{v}_a | \mathbf{x}_a, \tilde{\zeta}, \tilde{\delta})$. We have

$$\begin{aligned} E\{(\hat{\mu}_a - \hat{\mu}_a)(\hat{\mu}_a - \hat{\mu}_a)^T | \mathbf{y}_a\} &\approx E\{(\hat{\mu}_a - \hat{\mu}_a)(\hat{\mu}_a - \hat{\mu}_a)^T\} \\ &= E\{E((\hat{\mu}_a - \hat{\mu}_a)(\hat{\mu}_a - \hat{\mu}_a)^T | \mathbf{x})\} \\ &\quad + E\{E((\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T | \mathbf{x})\} \\ &= E\{(\hat{\mu}_a - \hat{\mu}_a)(\hat{\mu}_a - \hat{\mu}_a)^T\} \\ &\quad + E\{E((\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T | \mathbf{x})\}. \end{aligned}$$

The first approximation is correct to the order of $O_p(A^{-1})$, and can be justified as the number of areas tends to infinity. Intuitively, this makes sense if the information from any single area is asymptotically negligible compared to the information from all the other areas together. Next, the decomposition follows because $\tilde{\mu}_a - \hat{\mu}_a$ and $\hat{\mu}_a - \tilde{\mu}_a$ are independent of each other given \mathbf{x} : the former is a constant given \mathbf{x} .

In this way, we obtain an approximate CMSEP with a three-part decomposition

$$\text{CMSEP}_a \approx h_{1a}(\mathbf{x}_a; \zeta, \delta) + h_{2a}(\zeta, \delta) + h_{3a}(\mathbf{x}; \tilde{\zeta}, \tilde{\delta}, \psi)$$

where ψ contains the parameters of the conditional distribution of \mathbf{y}_a given \mathbf{x}_a , and

$$h_{1a}(\mathbf{x}_a; \zeta, \delta) \stackrel{\text{def.}}{=} B_a \text{Cov}(\mathbf{v}_a, \mathbf{v}_a | \mathbf{x}_a) B_a^T \quad (12)$$

$$h_{2a}(\zeta, \delta) \stackrel{\text{def.}}{=} E\{(\tilde{\mu}_a - \hat{\mu}_a)(\tilde{\mu}_a - \hat{\mu}_a)^T\} \quad (13)$$

$$h_{3a}(\mathbf{x}; \tilde{\zeta}, \tilde{\delta}, \psi) \stackrel{\text{def.}}{=} E\{(\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T | \mathbf{x}\}. \quad (14)$$

The three h -terms correspond, respectively, to a conditional prediction variance due to the random effects, a positive correction that accounts for the uncertainty in the estimation of the parameters based on the latent complete data, *i.e.*, the sampling variation, and another positive correction for the extra variation due to the randomness in the missing data. Alternative approximations are possible. For instance, one might use $E\{B_a \text{Cov}(\mathbf{v}_a, \mathbf{v}_a | \mathbf{x}_a) B_a^T | \mathbf{y}_a\}$ instead of h_{1a} , or replace h_{3a} with the unconditional $E\{(\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T\}$. The expressions (12) - (14) are chosen because they produce a clean separation between the sampling variation in the complete data and the extra variation owing to the missingness given the complete data. The difference from the CMSEP in the complete-data case (Booth and Hobert 1998) comes down to the third term h_{3a} .

4. Estimation

4.1 Parameter estimation

The structure of the data suggests an iterative procedure similar to the EM algorithm (Dempster, Laird and Rubin 1977). Given the current values of the parameters and the random effects, we calculate at the E-step the conditional expected complete two-way table $E(\mathbf{x} | \mathbf{y}, \mathbf{m})$. At the M-step we estimate the two random effects mixed models separately by some maximum penalized quasi-likelihood (MPQL) procedures. Iterations between the two yield an EMPQL algorithm.

For the E-step, let $I_{i,ak} = 1$ if the sample unit i belongs to the $(a, k)^{\text{th}}$ cell, and $I_{i,ak} = 0$ otherwise. The value is observed provided $r_{i,ak} = 1$, but is unknown if $r_{i,ak} = 0$. Let θ_{ak} be the generic compositions, depending of the adopted model. Suppose that

$$P[I_{i,ak} = 1 | i \in s] = d_{ak} \theta_{ak}$$

where s denotes the complete sample, and d_{ak} is some known constant which accounts for the eventual sampling design effect. For example, simple random sampling implies that $d_{ak} = 1$ for all (a, k) . An example of $d_{ak} \neq 1$ is when the sampling units are households, which are selected by a probability proportional to the household size. Let $m_{ak} = x_{ak} - y_{ak} = \sum_{i: r_{i,ak}=0} I_{i,ak} x_{i,ak}$. We have $E(x_{ak} | y_a, m_a) = y_{ak} + E(m_{ak} | m_a)$, where

$$\begin{aligned} E(m_{ak} | m_a) &= \sum_{i: r_{i,ak}=0} E(I_{i,ak} | r_{i,ak} = 0) x_{i,ak} \\ &= m_a P[I_{i,ak} = 1 | r_{i,ak} = 0] \\ &= m_a (1 - p_{ak}) d_{ak} \theta_{ak} / \left\{ \sum_j (1 - p_{aj}) d_{aj} \theta_{aj} \right\}. \end{aligned} \quad (15)$$

Having thus ‘completed’ the sample data, we move to the MPQL-step, where we apply the IWLS algorithm outlined in Section 2.1.3, respectively, to the complete-data model and the missing-data model conditional on the complete data.

4.2 Estimation of CMSEP

Evaluating the CMSEP at the estimated parameter values yields a plug-in estimate of the CMSEP. Of the three h -terms, h_{1a} is of the order $O_p(1)$, whereas both h_{2a} and h_{3a} are of the order $O_p(A^{-1})$, when the number of areas tends to infinity while the within-area sample sizes remain bounded. The results of Booth and Hobert (1998) and Prasad and Rao (1990), obtained in the univariate complete-data case, suggest that the bias in the plug-in estimate \hat{h}_{1a} is of the same order as \hat{h}_{2a} and \hat{h}_{3a} . These authors developed second-order correction through the Taylor expansion. We do not pursue such second-order asymptotics in this paper. Approximate expressions of the h -terms that accompany the EMPQL algorithm are given below.

Take first h_{1a} by (12). Based on the linearized data (6), the covariance matrix $\text{Cov}(\mathbf{v}_a, \mathbf{v}_a | \mathbf{z}_a)$ does not depend on either \mathbf{z}_a or \mathbf{x}_a . This is convenient because we then have

$$\begin{aligned} h_{1a}(\mathbf{x}_a; \zeta, \delta) &\approx B_a \text{Cov}(\mathbf{v}_a, \mathbf{v}_a | \mathbf{z}_a) B_a^T \\ &= B_a (G - G B_a^T V_a^{-1} B_a G) B_a^T \end{aligned} \quad (16)$$

where $V_a = B_a G B_a^T + Q R_a Q^T$ is the marginal covariance matrix of \mathbf{z}_a .

Next, take h_{2a} by (13). Let $\phi = (\zeta^T, \delta^T)^T$. Expanding $\tilde{\phi}$ around ϕ yields $\tilde{\mu}_a - \mu_a \approx \dot{\mu}_a' (\tilde{\phi} - \phi)$, where $\dot{\mu}_a' = \partial \mu_a / \partial \phi$, such that

$$h_{2a} \approx \dot{\mu}_a' \text{Cov}(\tilde{\phi}, \tilde{\phi}) \dot{\mu}_a'^T. \quad (17)$$

Based on (6) we derive $\dot{\mu}_a = H_a \zeta + D_a \dot{\mathbf{u}}_a$, where $D_a = B_a G B_a^T V_a^{-1}$ and $\dot{\mathbf{u}}_a = \mathbf{z}_a - H_a \zeta$. Denote by I the identity matrix. The partial derivatives in $\dot{\mu}_a'$ are given by

$$\partial \dot{\mu}_a' / \partial \zeta = (I - D_a) H_a$$

and

$$\partial \dot{\mu}_a' / \partial \delta_j = (\partial D_a / \partial \delta_j) \dot{\mathbf{u}}_a = (I - D_a) B_a (\partial G / \partial \delta_j) B_a^T V_a^{-1} \dot{\mathbf{u}}_a$$

where δ_j is the j^{th} variance parameter in the covariance matrix $G(\delta)$ of \mathbf{v}_a . To obtain $\text{Cov}(\tilde{\phi}, \tilde{\phi})$, suppose that the PQL approach is based on the following quasi log-likelihood

$$\ell = \sum_a \ell_a$$

and

$$\ell_a = -\frac{1}{2} \log |V_a| - \frac{1}{2} (\mathbf{z}_a - H_a \zeta)^T V_a^{-1} (\mathbf{z}_a - H_a \zeta).$$

The so-called sandwich formula yields then

$$\text{Cov}(\tilde{\phi}, \tilde{\phi}) = \left(-\frac{\partial^2 \ell}{\partial \phi^2} \right)^{-1} \left\{ \sum_{a=1}^A \left(\frac{\partial \ell_a}{\partial \phi} \right) \left(\frac{\partial \ell_a}{\partial \phi} \right)^T \right\} \left(-\frac{\partial^2 \ell}{\partial \phi^2} \right)^{-1}.$$

Finally, take h_{3a} by (14). Similarly as above we have $\tilde{\mu}_a = (I - \tilde{D}_a) H_a \zeta + \tilde{D}_a \tilde{\mathbf{z}}_a$ evaluated at $\phi = \tilde{\phi}$, and $\dot{\mu}_a = (I - \hat{D}_a) H_a \zeta + \hat{D}_a \hat{\mathbf{z}}_a$, where $\hat{\mathbf{z}}_a$ is derived from $\hat{\mathbf{t}}_a = \mathbf{t}(\hat{\mathbf{x}}_a)$ for $\hat{\mathbf{x}}_a = E(\mathbf{x}_a | \mathbf{y}_a, m_a; \hat{\phi}, \hat{\psi})$. Expanding $\hat{\phi}$ around $\tilde{\phi}$ and retain only the leading term, we obtain

$$\tilde{\mu}_a - \mu_a \approx \tilde{\mu}_a - \mu_a = \tilde{D}_a (\tilde{\mathbf{z}}_a - \mathbf{z}_a)$$

where $\tilde{\mu}_a = (I - \tilde{D}_a) H_a \zeta + \tilde{D}_a \tilde{\mathbf{z}}_a$, and $\tilde{\mathbf{z}}_a$ is derived from $\tilde{\mathbf{t}}_a = \mathbf{t}(\tilde{\mathbf{x}}_a)$ for $\tilde{\mathbf{x}}_a = E(\mathbf{x}_a | \mathbf{y}_a, m_a; \tilde{\phi}, \tilde{\psi})$. That is, we ignore the terms involving $\tilde{\phi} - \phi$. The remaining variation in $\tilde{\mathbf{z}}_a$ is due to the estimation of the missing-data model alone. Expanding $\tilde{\psi}$ around ψ , we obtain, by the chain rule,

$$\begin{aligned} h_{3a} &\approx C_a \text{Cov}(\hat{\psi}, \hat{\psi} | \mathbf{x}) C_a^T \\ \text{and} \quad C_a &= \left\{ D_a \left(\frac{\partial \mathbf{z}_a}{\partial \mathbf{t}_a} \right) \left(\frac{\partial \mathbf{t}_a}{\partial \mathbf{x}_a} \right) \left(\frac{\partial \mathbf{x}_a}{\partial \mathbf{p}_a} \right) \left(\frac{\partial \mathbf{p}_a}{\partial \eta_a} \right) \left(\frac{\partial \eta_a}{\partial \psi} \right) \right\}_{\phi=\tilde{\phi}, \psi} \end{aligned} \quad (18)$$

where we assume that $E(\hat{\psi} | \mathbf{x}) = \psi$ and $E[\tilde{\mathbf{z}}_a | \mathbf{x}] = \tilde{\mathbf{z}}_a$. Whereas the sandwich formula yields $\text{Cov}(\hat{\psi}, \hat{\psi} | \mathbf{x})$ under the conditional model of \mathbf{y} given \mathbf{x} , similarly to $\text{Cov}(\tilde{\phi}, \tilde{\phi})$ above.

4.3 Estimation of small area compositions

Suppose first that the GLSMM, defined by (2) and in combination with (5), has been estimated, upon which we obtain $\hat{\mu}_a^X$, and $\hat{\theta}_{ak}^X = \exp(\hat{\mu}_{ak}^X) / \sum_j \exp(\hat{\mu}_{aj}^X)$.

When the marginal totals $X_{.k}$ and $X_{.k}$ are known, it makes sense to apply the IPF, starting with the estimated table $\{\hat{\theta}_{ak}^X\}$. The difference from SPREE, which starts with the auxiliary table \mathbf{X}^0 , is that the interactions have been re-estimated. On convergence we obtain the estimated small area counts, denoted by $\hat{\mathbf{X}} = \{\hat{X}_{ak}^X\}$, and the corresponding compositions, denoted by $\hat{\theta}_{ak}^X = \hat{X}_{ak}^X / \sum_j \hat{X}_{aj}^X$, which are different from the direct model estimates $\hat{\theta}_{ak}^X$ that have provided the starting values for the IPF.

Often in practice, while the area totals $\{X_{.k}\}$ may be known, the marginal totals $\{X_{.k}\}$ need to be estimated based on the survey data available, separately using a method that is appropriate for the aggregated level. The IPF is still worth considering as long as these estimated marginal totals are judged to be more reliable and/or less biased than the aggregated small area estimates $\sum_a X_{ak} \hat{\theta}_{ak}^X$. The reason is that the estimated interactions $\hat{\alpha}_{ak}^X$ are preserved in the IPF, i.e., $\alpha_{ak}^X = \hat{\alpha}_{ak}^X$. By the log-linear identity (3), the difference between the direct model estimate $\hat{\theta}_{ak}^X$ and final estimate $\hat{\theta}_{ak}^X$ is due to the difference in the estimates of the main effects $\{\alpha_k^X\}$. Thus, less biased estimates of $\{X_{.k}\}$ are expected to yield less biased estimates of $\{\alpha_k^X\}$ and, thereby, less biased estimates of $\{\theta_{ak}^X\}$.

Suppose next that the MSLMM (7) combined with (5) have been estimated. We may express the interest of estimation, i.e., $\{\mu_{ak}^X\}$, in terms of \mathbf{z}_a defined as

$$\mathbf{z}_a = H_a \zeta + B_a \mathbf{v}_a + \mathbf{e}_a = H_a \zeta + B_a \mathbf{v}_a + \mathbf{e}_a^X + \mathbf{e}_a^{xlX}$$

$$= \mu_a^X + \mathbf{e}_a^{xlX} = H_a \zeta + \mathbf{v}_a^X + \mathbf{e}_a^X$$

where $\mathbf{e}_a^X = Q(\theta_a^X - \theta_a)$ and $\mathbf{e}_a^{xlX} = Q(\mathbf{t}_a - \theta_a^X)$. In accordance we have $R_a^X = R_a^X + R_a^{xlX}$, where $R_a^X = \text{Cov}(\theta_a^X, \theta_a^X | \theta_a)$ and $R_a^{xlX} = \text{Cov}(\mathbf{t}_a, \mathbf{t}_a | \theta_a^X)$. It follows that

$$\hat{\mu}_a^X = H_a \hat{\zeta} + (B_a \hat{G} B_a^T + \hat{Q} \hat{R}_a^X \hat{Q}^T) \hat{\mathbf{v}}_a^{-1} (\hat{\mathbf{z}}_a - H_a \hat{\zeta}). \quad (19)$$

The rest follows as above where μ_a^X is estimated directly under the GLSMM.

5. Example: Register-based small area household compositions

5.1 Register household data

Register-based household data have undergone considerable development in Norway. One of the goals is to produce detailed household statistics that traditionally are

only available from the census. For this purpose the registration of a unique dwelling identity number (DIN) was initiated in the last census in 2001. The work is not yet completed, and the DIN is still missing for about 6% of the people residing in the country. The rate of missing is differential as it varies over the household type as well as across the Municipalities, the latter of which is a reflection of the overall effort of the local administration regarding the registration of the DINs.

A household register can be compiled in a year after the census based on a number of data sources. The most important ones include the central population register (CPR), the DIN-register and the census household file (CH01). Even without the DIN a register household can be compiled based on the other information available. But the result suffers from informative under-registration of the DIN. For instance, a typical source of bias is cohabitants living without children, because such a couple appear as two single-person households in the CPR, unless they have already been identified as a household in the CH01. Nevertheless, historic as well as cross-country comparisons suggest that the national totals are acceptable. A more urgent problem lies on lower levels of aggregation. For example, changes from the census in 2001 are unlikely large in certain Municipalities, including the capital city Oslo where the increase in the proportion of single-person households is almost three times as high as it is in the rest of the country - see top-left plot in Figure 1. And a large part of the problem in Oslo can be explained by a combination of high proportion of cohabitants living without children and low DIN-registration rate (indeed, the lowest in the country).

5.2 Set-up of data

We shall illustrate our approach using these register household data. The target population contains all persons living at multiple-dwelling addresses at the beginning of year 2005, who do not belong to households of married people or registered partners; the latter household types are excluded because the DIN is not critical for compiling the households of these people. There is no distinction between the finite population and the sample in this case, i.e., $\mathbf{X} = \mathbf{x}$. The households that have registered DINs are treated as the 'observed' sample \mathbf{y} , whereas the households that do not have registered DINs are viewed as the missing. In this way the population consists of 713,387 persons, of which 558,136 persons have registered DINs. The overall rate of missing is about 22%.

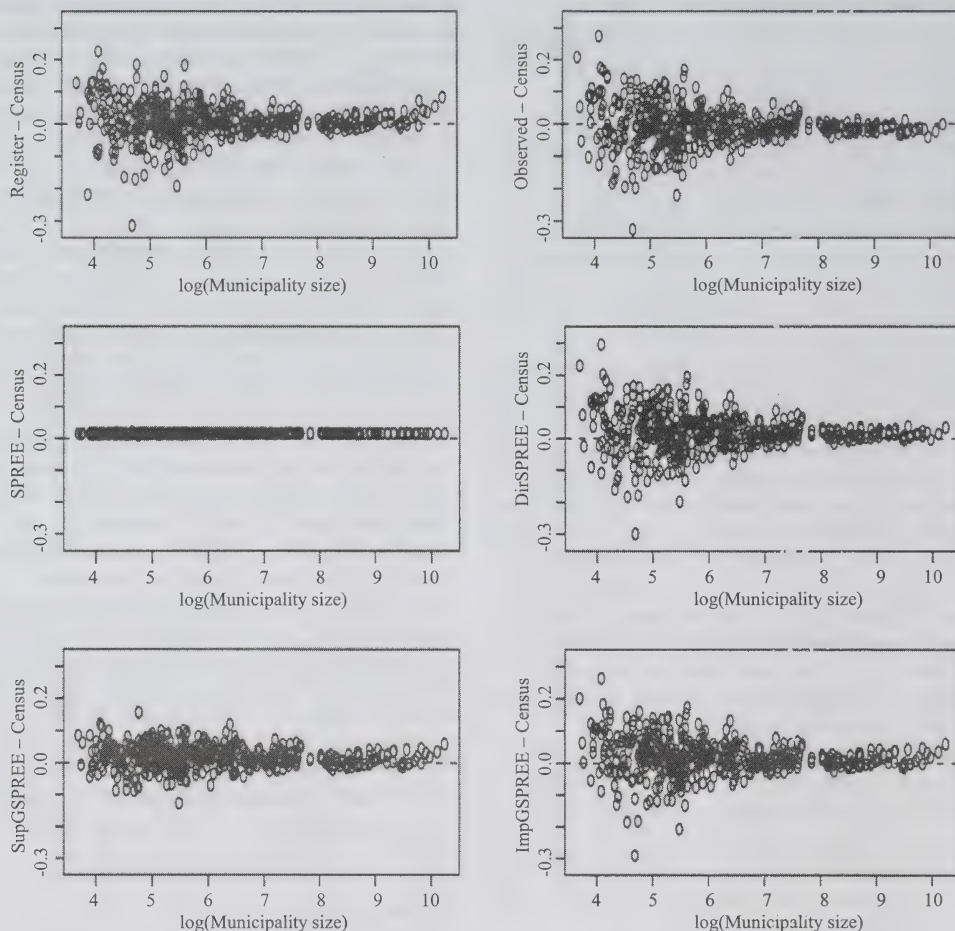


Figure 1 Difference between estimates of proportion of Single-person households and census counts in 2001 against log Municipality size: Register households (top-left), Households with registered DINs (top-right), SPREE based on census (middle-left), DirSPREE based on households with registered DINs (middle-right), SupGSPREE of super-population proportions (bottom-left), and ImpGSPREE of imputed finite-population proportions (bottom-right). The dashed line marks no difference

Let the Municipalities be the small areas of this study, where $A=433$. The households are classified into 4 categories: $k=1$ for “Single-person”, $k=2$ for “Single-parent”, $k=3$ for “Cohabitants”, and $k=4$ for “Other”, *i.e.*, $K=4$. Let i index the households, and let x_i be the number of persons living in the household. Let $X_{ak} = x_{ak}$ be the number of persons in the $(a, k)^{\text{th}}$ cell in the population, and let y_{ak} be the corresponding ‘observed’ cell count. Let N_{ak} be the number of households in the $(a, k)^{\text{th}}$ cell, and let n_{ak} be the corresponding number of ‘observed’ households. Notice that only the total number of persons is

known in each area, but not the total number of households. However, provided cell-specific probability of DIN-registrations, an estimator of N_{ak} based on \hat{X}_{ak} is given by $\hat{N}_{ak} = n_{ak} \hat{X}_{ak} / y_{ak}$. We shall therefore concentrate on the estimation of X_{ak} here.

Let $\{X_{ak}^0\}$ be the corresponding cell counts from the last census in 2001. Let $X'_{ak} = y_{ak} + m'_{ak}$ be the register counts in 2005, where m'_{ak} is the number of persons without the DIN. A register household can be considered as a form of imputed household that may suffer from informative missing of DINs. The register area total is correct, *i.e.*,

$X'_{.a} = X_{.a}$, and the national totals $\{X'_{.k}\}$ are considered acceptable. The question is whether estimates of $\{X'_{.ak}\}$ can be derived, based on the 'observed' \mathbf{y} and the allocation structure $\{X_{.a}\}$ and $\{X'_{.k}\}$, that better accounts for the differential missing DINs.

5.3 Set-up of model

Scatter plots of the register first-order interactions $\{\alpha_{.ak}^{X'}\}$ against the census interactions $\{\alpha_{.ak}^0\}$ provide motivation for the PIMM (4). To choose between the GLSMM (2) and the MSLMM (7), we look at the difference between the register proportion $\theta_{.ak}^{X'}$ and the corresponding census proportion $\theta_{.ak}^0$, i.e., $\theta_{.ak}^{X'} - \theta_{.ak}^0$, plotted against $\log X_{.a}$: the case of $k=1$ is shown in the top-left plot of Figure 1. Clearly, the variance of the difference increases as $X_{.a}$ decreases, and is not constant of $X_{.a}$. Notice that we are dealing with estimation at a very low level of aggregation here, where e.g., the median value of all $\{X'_{.ak}\}$ is only 70. We therefore adopt the model (7) for $\theta_{.ak}$, the quasi-likelihood (5) for $X_{.ak} = x_{.ak}$, and the quasi-likelihood (8) and the model (9) for $y_{.ak}$.

For the quasi-likelihood (5) we assume $v_1 = 1$. Let $t_{.ak} = X_{.ak}/X_{.a}$. We have

$$V(t_{.ak}) = N_a^{-1} \theta_{.ak} (1 - \theta_{.ak}) \bar{X}_a^{(2)} / \bar{X}_a^2$$

and

$$\text{Cov}(t_{.ak}, t_{.aj}) = -N_a^{-1} \theta_{.ak} \theta_{.aj} \bar{X}_a^{(2)} / \bar{X}_a^2.$$

where $\bar{X}_a^{(2)} = \sum_{i=1}^{N_a} x_i^2 / N_a$ and $\bar{X}_a = X_{.a} / N_a$. Since $x_i \geq 1$, we have $\bar{X}_a^{(2)} \geq \bar{X}_a^2$, and over-dispersion compared to the Multinomial- (N_a, θ_a) distribution. We calculate the factor $\bar{X}_a^{(2)} / \bar{X}_a^2$ based on the register data, which is then used as $\bar{X}_a^{(2)} / \bar{X}_a^2$ in the estimation below. Moreover, for the quasi-likelihood (8) we assume $v_2 = 1$, and

$$\begin{aligned} V(y_{.ak} | n_{.ak}) &= V\left(\sum_{i=1}^{n_{.ak}} r_{i, .ak} x_{i, .ak}\right) \\ &= \left(\sum_i x_{i, .ak}^2\right) V(r_{i, .ak}) \Rightarrow c_{.ak} = \left(\sum_i x_{i, .ak}^2\right). \end{aligned}$$

5.4 Estimation results

Six different estimators of the proportion of Single-person households (i.e., for $k=1$) are illustrated in Figure 1.

To start with, we have the direct register proportions $\theta_{.a1}^{X'}$ in the top-left plot, and the 'observed' proportions $\theta_{.a1}^y$ in the top-right plot. On average the proportion is increased based on the entire register compared to the census in 2001, whereas it is slightly decreased according to the 'observed' part only. This demonstrates that the missing DINs are informative, as explained before. Inclusion of the register households without the DINs raises the proportion of Single-person households. But the result is implausible in some of the largest Municipalities. Of course, large bias also

exists among the smaller Municipalities, but these are not easily detectable in a plot like this one.

Next, in the middle-left plot of Figure 1, estimates are obtained by SPREE using the census counts $\{X_{.ak}^0\}$ as the starting values. For the simple two-way table here, this yields an almost constant adjustment of the census proportions, with negligible change in the between-area variation. In the middle-right plot, estimates are obtained by SPREE using the 'observed' table $\{y_{.ak}\}$ as the starting values. Notice that, to start with the observed sample counts would be too unstable to be useful in usual survey sampling situations, but it is a viable option here because of the large amount of 'observed' data. To distinguish from the standard SPREE we shall refer to it as the *direct* SPREE (DirSPREE). As noted earlier, DirSPREE is unbiased under the assumption (10) of informative missingness. Indeed, it is seen to lead to useful adjustments for the largest Municipalities.

In the bottom-row plots of Figure 1, estimates are obtained using the double-mixed modeling approach. The estimates of the bottom-left plot are obtained by the IPF starting with the estimated super-population compositions $\{\hat{\theta}_{.ak}\}$, denoted by *SupGSPREE*. The extreme post-censal development in the largest Municipalities are reduced. But the changes from the census-proportions are clearly over-shrunk towards to the population average for the smaller areas. The variation is e.g., much less than that of $\theta_{.a1}^{X'} - \theta_{.a1}^0$ in the top-left plot. The estimates of the bottom-right plot are derived from the imputed finite-population counts, denoted by *ImpGSPREE*, which are calculated at the E-step of the EMPQL algorithm. The estimates for the largest Municipalities are similar to those of *SupGSPREE*, and the variation in the changes from the census-proportions is similar to that of DirSPREE.

5.5 Estimation of CMSEP

Approximate CMSEP of the ImpGSPREE compositions can be derived similarly as in Section 3. Denote by $\hat{X}_{.ak}$ the ImpGSPREE count, and by $\check{X}_{.ak}$ the BP based on known conditional distribution of \mathbf{X}_a given (\mathbf{y}_a, m_a) . We have

$$\begin{aligned} \text{CMSEP}(\hat{\mathbf{X}}_a) &\approx E\{(\hat{\mathbf{X}}_a - \mathbf{X}_a)(\hat{\mathbf{X}}_a - \mathbf{X}_a)^T | \mathbf{y}_a, m_a\} \\ &\quad + E\{(\hat{\mathbf{X}}_a - \check{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \check{\mathbf{X}}_a)^T\}. \end{aligned}$$

Moreover, let $\tilde{\phi}$ be the hypothetical estimate of ϕ based on the complete data $\mathbf{x} = \mathbf{X}$, and let $\hat{\psi}$ be the estimate of ψ based on the observed data. Let \mathcal{Q}_1 and \mathcal{Q}_2 be, respectively, the Jacobian matrix of partial derivatives $\partial \hat{\mathbf{X}}_a / \partial \phi$ and $\partial \hat{\mathbf{X}}_a / \partial \psi$. We have

$$\begin{aligned} E\{(\hat{\mathbf{X}}_a - \check{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \check{\mathbf{X}}_a)^T\} \\ \approx E\{(\tilde{\mathbf{X}}_a - \check{\mathbf{X}}_a)(\tilde{\mathbf{X}}_a - \check{\mathbf{X}}_a)^T\} \\ + E\{(\hat{\mathbf{X}}_a - \tilde{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \tilde{\mathbf{X}}_a)^T | \mathbf{X}\} \\ \approx \mathcal{Q}_1 \text{Cov}(\tilde{\phi}, \tilde{\phi}) \mathcal{Q}_1^T + \mathcal{Q}_2 \text{Cov}(\hat{\psi}, \hat{\psi} | \mathbf{X}) \mathcal{Q}_2^T. \end{aligned}$$

Together, these lead to a three-part decomposition of the CMSEP similar to (12) - (14). In the estimation of the CMSEP below we ignore the effect of IPF. This is justified in our case because the IPF essentially amounts to a constant multiplicative adjustment very close to unity, as can be seen in the middle-left plot in Figure 1.

The CMSEP of a DirSPREE count is calculated as a 'sampling' variance that is induced by missing-at-random within each cell of the two-way table, plus a squared bias term which is estimated by the squared difference between the ImpGSPREE count and the corresponding DirSPREE count, provided the assumption (9) is a more appropriate model for the missing data than the assumption (10).

The estimated root CMSEPs (rmsep) are given in Figure 2. On average both are decreasing as the Municipality size

increases. However, for some of the largest Municipalities, the CMSEP of the DirSPREE proportion is abnormally large for Single-person and Cohabitants households due to the bias term. On the whole the CMSEP of the ImpGSPREE composition is clearly smaller than that of the DirSPREE. The h_{1a} -term, corresponding to the prediction variance of \mathbf{X}_a , is by far the dominating contribution to the CMSEP (over 99% in many areas). This is understandable since there are over 550 thousand people in the 'observed' sample, such that the uncertainty in parameter estimation is comparatively negligible. But the quoted percentage will be lower in a sample survey situation, as the estimation uncertainty summarized in terms h_{2a} and h_{3a} increases.

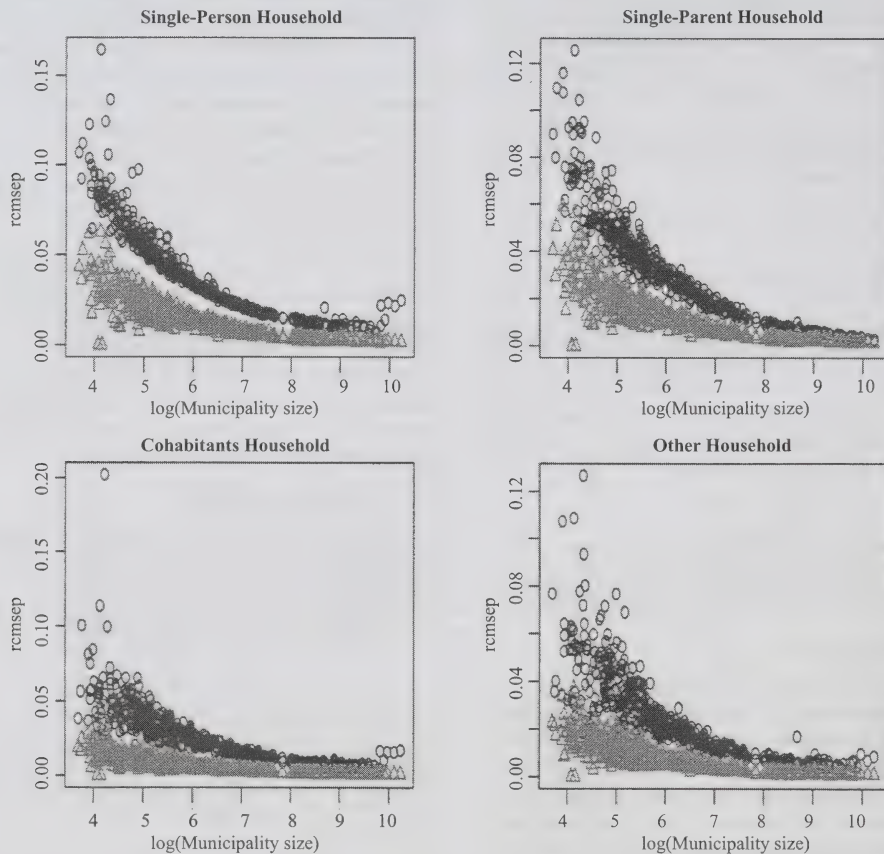


Figure 2 Estimated root conditional mean squared error of prediction (rmsep) of DirSPREE (circle) and ImpGSPREE (triangle) of Municipality household proportions

6. Summary

In the above we outlined a double-mixed modeling approach that extends the GSPREE methodology to estimation of small area compositions subjected to differential missing data. An approximate CMSEP was derived which contains a three-part decomposition, corresponding to the prediction variance of the unknown random effect, the sampling variance in the absence of missing data, and the extra variance due to the missing data, respectively. The approach was applied to the Norwegian register household data, which yielded useful adjustments for informative missing of dwelling identity numbers.

Acknowledgements

I am thankful to the referees and the Associate Editor for comments and suggestions that have helped to improve the presentation.

References

- Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.
- Booth, J.G., and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.
- Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 93-273-282.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Longford, N. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society, Series A*, 162, 227-245.
- McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- Prasad, N.G., and Rao, J.N.K. (1990). The estimation of mean square errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.
- Zhang, L.-C., and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 479-496.

Small area population prediction via hierarchical models

Debora F. Souza, Fernando A.S. Moura and Helio S. Migon¹

Abstract

This paper proposes an approach for small area prediction based on data obtained from periodic surveys and censuses. We apply our approach to obtain population predictions for the municipalities not sampled in the Brazilian annual Household Survey (PNAD), as well as to increase the precision of the design-based estimates obtained for the sampled municipalities. In addition to the data provided by the PNAD, we use census demographic data from 1991 and 2000, as well as a complete population count conducted in 1996. Hierarchically non-structured and spatially structured growth models that gain strength from all the sampled municipalities are proposed and compared.

Key Words: Markov Chain Monte Carlo (MCMC); Population projection; Spatial models.

1. Introduction

Like many other countries, the demand for detailed and updated small area statistics has been steadily growing in Brazil. This increasing demand is motivated by the need to have a more precise picture of subregions and has been driven by issues of distribution, equity and disparity. For instance, there may exist subregions or subgroups that are not keeping up with the overall average in certain respects. Therefore, there is a need to identify such regions and to have statistical information at that geographical level before taking any possible remedial action. Besides these national requirements, local authorities are faced with the need of having reliable estimates, such as demographic characteristics, for analysis, planning and administration purposes.

In Brazil, one important example of the demand for reliable estimates is related to how constitutionally mandated federal revenue sharing is apportioned annually to the various municipalities (Brazil is a federated republic made up of states and the Federal District. The states are divided into municipalities, which share characteristics of cities and counties - they can contain more than one urban area, but they have a single mayor and municipal council). The predicted number of inhabitants in a municipality is used by the federal government as a criterion to distribute funding. Hence, there is a need to obtain reliable municipal population forecasts in order to fairly apply this criterion, regulated by federal law.

An important source of demographic data is the annual Household Survey (PNAD). However, this survey is not designed to produce estimates at the municipal level. In other words, apart from a few municipalities, the municipal sample sizes are not large enough to yield acceptable standard errors when the direct survey estimates are used.

Furthermore, a considerable number of municipalities are not sampled at all.

The current approach to obtain municipal population estimates is based on making prediction for a larger area at first, and then using some auxiliary information to allocate the total predicted population to the municipalities. In turn, prediction for a larger area is done by assuming that birth, mortality and migration rates are the same for all municipalities. The major drawback of this approach is that it relies on the assumed model evolution. It does not take into account all uncertainties and does not provide, in general, error measures of the estimates.

The small area estimation problem has received attention in the statistical literature due to the growing demand for detailed statistical information from the public and private sectors. An excellent and updated account of methods and applications of small area estimation can be found in Rao (2003). The main source of small area data is provided by periodic surveys whose sample sizes are not large enough to provide reliable estimates for the areas. One way of tackling this problem is to gain strength from all areas and through other sources of related data. As stated in Pfeffermann (2002), the sources of data suitable for this task can be classified into two categories: data obtained from other similar areas with respect to the characteristic of interest and past data obtained for the characteristic of interest and auxiliary information. In our demographic context, the main source of related data is provided by the 1991 and 2000 censuses and a complete count of the population carried out in 1996.

The aim of this work is to obtain estimates of the municipal populations based on survey data provided by the PNAD and census data. A non-structured hierarchical model is proposed and its fitness and predictive power are

1. Debora F. Souza, Department of Methods and Quality, IBGE, Rio de Janeiro, 20031-170. E-mail: debora.souza@ibge.gov.br; Fernando A.S. Moura, Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil. E-mail: fmoura@im.ufrj.br; Helio S. Migon, Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil. E-mail: migon@im.ufrj.br.

evaluated. We also consider a spatially structured hierarchical model, in the spirit of Moura and Migon (2002), since the population per area and its growth pattern might be related to the development of its neighboring areas. For the sake of simplicity, from now on we respectively call the non-structured hierarchical and spatially structured hierarchical models as the Hierarchical model and Spatial model.

In Section 2 the main data sources used in this work are described. In Section 3, the proposed models and a model selection criteria are presented. Applications with real and a simulated data are presented in Section 4. Finally, Section 5 contains a brief summary with an outline for future research.

2. Data set

The input data for the models introduced in Section 3 are taken from the annual Household Surveys (PNADs) from 1992 to 1999, the 1991 and 2000 census data and a complete enumeration of the population carried out in 1996. In order to evaluate the proposed approach, the municipalities of São Paulo State are considered as the areas of interest.

In this section we present a brief description of these data sources, reporting their main advantages and limitations. The population direct estimates of sampled municipalities were obtained from the PNAD. As explained in Section 3, these estimates are regarded as the input data for making inference about our target parameters. The two censuses and the 1996 population count are also utilized in our application.

The Brazilian Demographic Census is the main source of information about the population. It is carried out every ten years, usually in the beginning of the decade. Although the objective is to count all the population, some enumeration errors are found. The magnitude of the errors is evaluated through a post enumeration survey carried out soon after the completion of the census.

The annual Household Survey (PNAD) is designed to produce basic information about the socioeconomic situation of the country. The investigation unit is the household, for which yearly information about the number of dwellers, their gender, education level, employment, *etc.* is collected. The survey is not carried out in a census year, and was also not conducted in 1994 for administrative reasons. The sample is selected by a three-stage cluster sampling design. The primary and secondary units are respectively the municipality and enumeration areas (with 250 households on average). The municipalities are stratified according to their population sizes as obtained from the last census. In the first stage, all municipalities belonging to the metropolitan regions and the state capitals (which in Brazil are normally the largest cities in the respective states) are sampled. The municipalities whose

populations are greater than some cutoff value are also included in the sample with probability one. The ones left are stratified and two of them are sampled from each stratum with probability proportional to their population sizes.

The enumeration areas are sampled with probability proportional to the number of households residing in the area in the last census. Finally, in the last stage the households are sampled systematically with equal probability from a list, which is updated at the beginning of the survey. The municipalities and enumeration districts are kept the same in all the surveys carried out in the same decade, while households are sampled every year.

Since each area is sampled with probability proportional to its respective number of households, it could be argued that the sampling mechanism is informative with respect to the population of the area. However, since the response variable actually used in this work is the area density, it is reasonable to assume that the sample selection mechanism is not relevant. Thus, this issue is not exploited in this work. A good reference about how to make small area inference under informative sampling is Pfeffermann and Sverchkov (2007). We also recommend Pfeffermann, Moura and Silva (2006) for readers interested in how to employ a Bayesian approach to hierarchically modeling under informative sampling.

3. Model specification

3.1 Exponential growth model

Let y_t be sample values of a distribution belonging to an exponential family with expected value given by $\pi_t = E(y_t | \theta_t)$ where θ_t is a vector of unknown parameters.

An important and wide class of exponential growth models parameterized by $(\alpha, \beta, \gamma, \phi)$ is defined as:

$$\pi_t = [\alpha + \beta \exp(\gamma t)]^{1/\phi}. \quad (1)$$

Some special well-known cases in the literature are:

- (1) Logistic: with $\phi = -1$, $\pi_t^{-1} = \alpha + \beta \exp(\gamma t)$;
- (2) Gompertz: with $\phi = 0$, defining (1) as $\log(\pi_t) = \alpha + \beta \exp(\gamma t)$;
- (3) Modified exponential: with $\phi = 1$, $\pi_t = \alpha + \beta \exp(\gamma t)$.

The main advantage of using model (1) is the possibility of keeping the observations y_t in the original scale, changing only the trajectory of π_t , making interpretation easy. Furthermore, the time intervals do not need to be of the same length, allowing the data to come from different reference sources (see Section 4 for further details).

When $\psi = \exp(\gamma) < 1$, the process is non-explosive, implying that π_t converges to $\alpha^{1/\phi}$ when $t \rightarrow \infty$, with the

convention that for $\phi = 0$, this quantity is equal to $\log(\alpha)$. When $\psi > 1$, the curves are concave for $\phi \geq 0$ and $\beta > 0$, leading to an explosive process. This class of models is called the generalized exponential growth model. Migon and Gamerman (1993) show how the exponential growth model can be viewed as a particular case of a general dynamic model.

3.2 Hierarchical growth models

In this paper our main parameters of interest π_{it} are nonlinear exponential growth functions with some parameters that are hierarchically or spatially structured. Spatially structured models provide alternative ways for connecting similar neighboring areas. We further assume that the sampling variance σ_{it}^2 follows a model that depends on the sample size in the respective municipality. In this work, hierarchical and spatial models are fitted and compared.

We assume that the population sizes are available for all the m municipalities of São Paulo State for the census years of 1991 and 2000, as well as the complete population count in 1996. From now on, we simply refer to them as the census data. In order to improve the hypothesis of exchangeability of the parameters describing the mean of the process, our response variables are set as the sampled municipal density estimates instead of the municipal population estimates. See also the end of Section 2 for further reasons for using the densities.

For each period, estimates of these quantities are available only for $k < m$ first-stage units municipalities of the PNAD sample. In order to estimate the municipal density, we simply divide the total population estimate by the respective municipal area.

Let y_{it} be the population density obtained from the census data or estimated by the PNAD at time t , $t = 1, \dots, n$ for the i^{th} municipality, $i = 1, \dots, m$. Our aim is to make inferences about the true population density π_{it} for the population of all municipalities, including those that are not sampled. In the next section, true municipal population densities π_{it} are modeled via a stochastic nonlinear hierarchical growth function. We assume that the random quantities y_{it} are normally distributed with mean π_{it} and variance σ_{it}^2 .

We use a Bayesian approach in this work. Therefore, predictions are described by probability distributions, giving the opportunity for users to analyze the uncertainties involved in the decision process. This fact is one of the advantages, among many others, of using this kind of approach.

Only in the census years are the y_{it} obtained for all the municipalities of São Paulo State. Although the census attempts to obtain complete enumeration of the whole population, coverage errors can occur. The following model

is assumed therefore for the census data and the data obtained from the PNAD, with exception that the variances σ_{it}^2 are set to be smaller for the census data (see Section 3.4 and also the final remarks in Section 5):

$$\begin{aligned} y_{it} &= \pi_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \sigma_{it}^2) \\ \pi_{it} &= \{\alpha_i + \beta \exp(\gamma_i t)\}^{1/\phi} \\ \alpha_i &= \alpha + \xi_{\alpha i}, \quad \xi_{\alpha i} \sim N(0, \sigma_{\xi_{\alpha}}^2) \\ \gamma_i &= \gamma + \xi_{\gamma i}, \quad \xi_{\gamma i} \sim N(0, \sigma_{\xi_{\gamma}}^2) \end{aligned} \quad (2)$$

where the prior distributions of α , β and γ are given by: $\alpha \sim N(\mu_{\alpha}, \sigma_{\alpha}^2)$, $\beta \sim N(\mu_{\beta}, \sigma_{\beta}^2)$, $\gamma \sim N(\mu_{\gamma}, \sigma_{\gamma}^2)$. It should be noted that information from all areas is obtained through the hierarchical structure of the parameters α_i and γ_i . Another way of borrowing information between municipalities is to assume that α_i are spatially structured (see Section 3.3). Supposing that the mean π_{it} is non-explosive, the parameter $\alpha^{1/\phi}$ can be regarded as the value at which the mean municipal population stabilizes. The parameters β and γ affect the evolution of the density over time. The prior distributions of α , β and γ can be chosen by taking advantage of some prior demographic knowledge of the expected population evolution. In our application, we set $\phi = 1$, implying that for $t = 0$ the true value density in each municipality is given by $\alpha_i + \beta$. The hierarchical structure imposed on the parameters α_i , implies that the expected value of the true density for any municipality at $t = 0$ is $\alpha + \beta$. To assume that the growth parameters, γ_i , have a hierarchical structure means that the densities have different growth rates but share the same mean. A small simulation study (see Section 4.1) guides us to keep the β parameter fixed for all areas, without any loss of generality, since the levels are still different for different municipalities. In all models considered in our application, we assume that $\tau_{\alpha}^2 = \sigma_{\alpha}^2 \sim G(a_{\alpha}, b_{\alpha})$, $\tau_{\gamma}^2 = \sigma_{\gamma}^2 \sim G(a_{\gamma}, b_{\gamma})$. In order to assign vague priors, in Section 4.2 we set small values for the parameters related to these precision prior distributions.

The assumption that the mean function π_{it} is given by an exponential growth curve allows adjusting for increasing or decreasing population density. The sources of data used have different reference data and are not equally spaced in time. In this case, the use of an exponential growth curve yields an extra advantage, since we can simply make a scale of time in order to conform with the different data sources, as explained in the application section 4.

3.3 Spatial model

In the Hierarchical model presented in the previous section, the information from all areas is combined in order to predict the population of a particular area. However, it is reasonable to assume that two or more neighboring municipalities have more similar demographic densities

than two other arbitrarily chosen ones. The regional structure is represented in the joint prior distribution of the random spatial effects. We consider that two areas are neighbors if they share a border.

In our proposed model, the demographic density in an area i at time t , π_{it} , is affected by its neighboring areas by adding random spatial effects δ_{α_i} to the parameters α_i , that is, $\alpha_i = \alpha + \delta_{\alpha_i}$, where α is a term representing the intercept. Therefore, α_i vary only with the spatial effect, representing a local effect, while the growth parameters γ_i 's are regarded as similar among all areas (overall effect).

The relationship between neighboring areas is defined in the prior distributions of δ_{α_i} . The prior joint distribution of $\delta_{\alpha} = (\delta_{\alpha_1}, \dots, \delta_{\alpha_m})'$ given the hyperparameter σ_{α}^2 , is defined as in Mollié (1996):

$$p(\delta_{\alpha} | \sigma_{\alpha}^2) \propto \frac{1}{\sigma_{\alpha}^{m/2}} \exp \left\{ -\frac{1}{2\sigma_{\alpha}^2} \sum_{i=1}^m \sum_{k \in \partial i} w_{ik} (\delta_{\alpha_i} - \delta_{\alpha_k})^2 \right\} \quad (3)$$

where w_{ik} are the weights associated with the regional structure. The weights were chosen such that $w_{ik} = 1$, if i and k are contiguous, and $w_{ik} = 0$, otherwise. The distribution of $\delta_{\alpha} | \sigma_{\alpha}^2$ is evidently improper, since we can add any constant to all of the δ_{α_i} and $p(\delta_{\alpha} | \sigma_{\alpha}^2)$ is not affected. Thus, we must impose a constraint to ensure that the model is identifiable. We set $\sum_{i=1}^m \delta_{\alpha_i} = 0$ and assign a uniform prior distribution on the whole real line to the intercept α . It is not difficult to see that this procedure leads to a proper $(m-1)$ dimensional likelihood density, see Besag and Kooperang (1995) for further details.

The prior conditional distribution of δ_{α_i} , given the effects δ_{α_k} of the remaining areas and the hyperparameter σ_{α}^2 , is normal with mean and variance given by:

$$E[\delta_{\alpha_i} | \delta_{\alpha_k}, k \in \partial i, \sigma_{\alpha}^2] = \bar{\delta}_{\alpha_i}$$
$$\text{Var}[\delta_{\alpha_i} | \delta_{\alpha_k}, k \in \partial i, \sigma_{\alpha}^2] = \frac{\sigma_{\alpha}^2}{w_{i+}}$$

where $\bar{\delta}_{\alpha_i}$ denotes the arithmetic mean of the δ_{α_k} for $k \in \partial i$ (the contiguous areas of i), and $w_{i+} = \sum_{k=1}^m w_{ik}$ is the number of neighboring municipalities of i .

Figure 1 shows the demographic densities of São Paulo municipalities in 1991. These municipalities tend to be concentrated geographically according to density classes. This suggests that the spatial model can be usefully applied.

3.4 Modeling the sampling variances

Since we use data from two different sources, it makes sense to assume that the sampling variances vary over time. Furthermore, we can also consider that the variances change with the areas.

For the years in which the data are provided by the PNAD, we assume the following model for the sampling variances:

$$\log(\sigma_{it}^2) = \eta_0 + \eta_1 \cdot (1/n_i) \quad (4)$$

with n_i representing the number of enumeration areas sampled in the i^{th} area. This model captures the expectation that the variance gets smaller as the sample size increases.



Figure 1 Population densities of São Paulo municipalities in 1991

For the years that the censuses were carried out, we assumed that σ_{it}^2 is known and $\log(\sigma_{it}^2) = \log(v_{it})$ where v_{it} is calculated in such a way that the census coverage error is 5% for all areas. This hypothesis implies that the true population in each area for census years lies in the interval given by the observed population in the census plus or minus 5% of this value. Therefore, for the census years we set the standard deviation as: $\sigma_{it} = 0.05 * (y_{it}/2)$. Assuming known variance in the census years is a way of giving more weight to census data, since one would expect a complete census to provide more reliable information than survey data. Independent normal distributions are assumed for the parameters η_0 and η_1 : $\eta_k \sim N(\mu_{\eta_k}, \phi_{\eta_k})$; $k = 0, 1$. In order to assign vague priors to the η 's, we set both prior means as zero and large values for the ϕ_{η} 's. See Section 4.2 for details.

3.5 Summary of the models

The prior distributions of the common parameters of the Spatial and Hierarchical models are the same as already described for the former. The distributions of the random spatial effects are specified in Section 3.3. The variance σ_{it}^2 in the Spatial model was stated as in the Hierarchical model. A summary of the models in Section 4 is presented in Table 1. For the sake of simplicity, the application was carried out by fixing $\phi = 1$ in both models.

3.6 Computational issues

The posterior distributions of the parameters for the models proposed cannot be obtained in closed forms. Therefore, it is necessary to use numerical approximation methods. One alternative, often used and easy to implement, is to generate samples of these distributions based on the Markov Chain Monte Carlo (MCMC) algorithm. Since the full conditional distributions of all the model parameters have closed form, except for the vector $\gamma = (\gamma_1, \dots, \gamma_k)$, we employed the Gibbs sampler algorithm with one acceptance/rejection algorithm step for sampling from the vector γ . Let π_{it} be the population density in the i^{th} area at time t . The following steps summarize how to sample from the posterior distribution of π_{it} :

1. Generate $\alpha_i^{(l)}, \beta^{(l)}, \gamma_i^{(l)}, \alpha^{(l)}, \gamma^{(l)}, \tau_{\alpha}^{2(l)}, \tau_{\gamma}^{2(l)}, \eta_0^{(l)}$ and $\eta_1^{(l)}$ for $l = 1, \dots, M$, where M is the number of MCMC samples generated from the full conditional distributions of all model parameters including the random effects;
2. Calculate $\pi_{it}^{(l)} = \alpha_i^{(l)} + \beta^{(l)} \exp(\gamma_i^{(l)} t)$;

Three informal checks for convergence, based on graphical techniques, were applied for assessing the convergence when fitting our proposed models. They consist of observing the histogram, the trace and the autocorrelation function for each of the sampled values calculated. The histogram analysis allows us to identify possible departures from convergence, such as the presence of multiple modes. The trace of the multiple chains simulated in parallel, each one with different starting points and overdispersed with respect to the target distribution, provides a rough indication of stationary behavior when the sequences of values tend to oscillate in the same region. The plot of the autocorrelation function allows identifying whether the sampling can be regarded as independent.

In addition to these informal checks, other more formal criteria were applied. The criteria introduced by Brooks and Gelman (1998) and implemented in WinBugs 1.4 (Spiegelhalter, Thomas, Best and Lunn 2004) permit diagnosing whether dispersion within chains is larger than dispersion between chains. Consider I parallel chain and a parameter of interest λ . Let λ_i^j be the j^{th} value of the i^{th} chain, for $i = 1, \dots, K$ and $j = 1, \dots, J$. Then the variances between chains \hat{B} and within chains \hat{W} are given by

$$\hat{B} = J(K-1)^{-1} \sum_{i=1}^K (\bar{\lambda}_i - \bar{\lambda})^2$$

and

$$\hat{W} = \{K(J-1)\}^{-1} \sum_{i=1}^K \sum_{j=1}^J (\lambda_i^j - \bar{\lambda}_i)^2$$

where $\bar{\lambda}_i$ and $\bar{\lambda}$ respectively are the average of observations of chain i , $i = 1, \dots, K$ and the global average. Under convergence, all these KJ values are drawn from the posterior of λ and the variance of λ can be consistently estimated by \hat{B} , \hat{W} and the weighted average $\hat{\delta}_{\lambda}^2 = (1 - 1/J) \hat{W} + (1/J) \hat{B}$.

Table 1
Summary of the models employed

model	parameters	variance	prior distribution
Hierarchical	$\alpha_i = \alpha + \xi_{\alpha_i}$	$\log(\sigma_{it}^2) = \eta_0 + \eta_1 (1/\pi_i)$	$\eta_0 \sim N(\mu_{\eta_0}, \phi_{\eta_0})$
	β	for survey data	$\eta_1 \sim N(\mu_{\eta_1}, \phi_{\eta_1})$
	$\gamma_i = \gamma + \xi_{\gamma_i}$	σ_{it}^2 is assumed to be known for census data	
Spatial	$\alpha_i = \alpha + \delta_{\alpha_i}$	$\log(\sigma_{it}^2) = \eta_0 + \eta_1 (1/\pi_i)$	$\delta_{\alpha_i} \delta_{\alpha_{-i}}, \tau_{\alpha}^2 \sim N(\bar{\delta}_{\alpha_i}, \tau_{\alpha}^2/w_{\alpha_i})$
	β	in the survey	$\sum_{i=1}^m \delta_{\alpha_i} = 0$
	$\gamma_i = \gamma + \xi_{\gamma_i}$	σ_{it}^2 is assumed to be known for census data	$\eta_0 \sim N(\mu_{\eta_0}, \phi_{\eta_0})$ $\eta_1 \sim N(\mu_{\eta_1}, \phi_{\eta_1})$

If the chains have not yet converged, then initial values will still be influencing the trajectories and $\hat{\sigma}_\lambda^2$ will overestimate σ_λ^2 until stationarity be reached. On the other hand, before convergence, \hat{W} will tend to underestimate σ_λ^2 . Following these reasoning, Brooks and Gelman (1998) proposed an iterated graphical approach, which is implemented in WinBugs 1.4. It allows to check if: (i) the weighted posterior variance estimated $\hat{\sigma}_\lambda^2$ and the within-chain variance \hat{W} stabilize as a function of J , and (ii) the variance reduction factor, $\hat{R} = \hat{\sigma}_\lambda^2/\hat{W}$, approaches 1.

4. Application

In this section we present two applications of our approach, the first one with a simulated data set and the second one with the real data set that motivated this work. The simulation study aims to check if the parameters of interest are being properly estimated, as well as to perform some sensitivity analysis with respect to the form of the prior distributions used for fitting the model.

4.1 Application to simulated data

We carried out a small simulation study fitting the Hierarchical and Spatial models presented in Section 3. The true model hyperparameters related to the growth curve were fixed as $\alpha = 40, \beta = 25, \gamma = 0.05$. Thus, we are considering a situation where the population size approximately doubles in 25 years. The parameters related to the sampling variance model were fixed as $\eta_0 = 6.5, \eta_1 = 0.5$. Finally, the precision parameters were respectively set as $\tau_\alpha^2 = 0.0001$ and $\tau_\gamma^2 = 400$. The precision τ_α^2 and τ_γ^2 were fixed to be in agreement with the scales of the quantities they respectively measure. The intercept presents more relative variation between areas than the growth parameter, which is expected in practical situations.

Since it is well recognized that the form of the priors has more impact on the component of variance parameters than the fixed parameters, we fitted the simulated data using two different vague priors for the parameters related to the variances: uniform for the standard deviation, which is one of the priors recommended by Gelman (2006) for linear hierarchical models, and gamma for the precision, commonly used as the default in some computational packages. In the first case, we assigned $\sigma_\alpha \sim U(0, 1,000)$ and $\sigma_\gamma \sim U(0, 100)$, where $\sigma_\alpha = 1/\tau_\alpha$ and $\sigma_\gamma = 1/\tau_\gamma$. In the second case, we considered $\tau_\alpha^2 \sim G(0.001, 0.001)$ and $\tau_\gamma^2 \sim G(0.001, 0.001)$. For the other parameters, we set $\alpha \sim U(-\infty, +\infty)$, for the Spatial Model (see Section 3.3 for further details) and $\alpha \sim N(0, 10^6)$ for the Hierarchical model. For the others parameters we set $\beta \sim N(0, 10^6)$, $\gamma \sim N(0, 10^2)$, $\eta_0 \sim N(0, 10^4)$ and $\eta_1 \sim N(0, 10^4)$ for both models. The effect of the number of small areas is also investigated. We simulated separate data from the Hierarchical and Spatial models with $m = 60$ and $m = 100$

areas in each case. For each combination of the number of areas and the model employed we generated 200 data sets. Therefore, a total of 800 sets of artificial data was simulated. The distribution of the sample sizes within the areas is the same for the simulated data sets with 60 and 100 areas. Table 2 presents the relative frequencies of the small areas sample sizes for the both simulated data sets. These sample sizes are very similar to the sample sizes in the real data that underlines this simulation study. The number of neighbors employed in the spatial model varies from 1 to 12 and each area has on average 5 neighbors. We considered a total period of $n = 9$ years.

Table 2
Relative frequencies of the small area samples sizes for both simulated data sets

Sample size	Relative frequency
2	0.05
5	0.20
8	0.25
10	0.25
12	0.20
15	0.05

In order to get rid of chain correlation, we generated 20,000 samples after discarding the first 10,000. There is no evidence for non-convergence of the Hierarchical and the Spatial model parameters. A careful analysis of some outputs obtained from the MCMC samples for some simulation sets suggests that convergence was achieved for all model parameters. We assessed the statistical properties of the population density (π_{it}) estimates by investigating the average of the absolute relative error of the estimates (ARE) and the mean square error (MSE), respectively given by:

ARE_{i,t} = \frac{1}{200} \sum_{l=1}^{200} \frac{|\hat{\pi}_{i,t}^{(l)} - \pi_{i,t}^{(l)}|}{\pi_{i,t}^{(l)}}

and

MSE_{i,t} = \frac{1}{200} \sum_{l=1}^{200} (\hat{\pi}_{i,t}^{(l)} - \pi_{i,t}^{(l)})^2,

$i = 1, \dots, m, t = 1, \dots, n$. There is no much variation, as far as the ARE values are concerned. For the two models fitted and both small area sample sizes tried, the ARE values are around 1.5%.

Table 3 shows a summary of the MSE values obtained from the simulations carried out under the Spatial and Hierarchical models with 60 and 100 areas and respectively assigning gamma and uniform priors to the precision and to the standard deviation of the parameters related to the variance. It can be seen from Table 3 that the MSEs are not affected by the use of different vague priors. It is noteworthy that increasing the number of areas from 60 to 100 results in a small decrease of 6% in the median of the MSE for the Spatial model. However, for the case of the Hierarchical model, the decrease is about 13%.

Table 3
Summary of mean square error distribution for the spatial and hierarchical models

Model	Num. of areas	Gamma prior			Uniform prior		
		1 st Qu.	Median	3 rd Qu.	1 st Qu.	Median	3 rd Qu.
Spatial	60	0.398	1.741	3.574	0.394	1.737	3.595
	100	0.525	1.637	3.538	0.524	1.641	3.517
Hierarchical	60	0.542	2.218	6.262	0.646	2.223	6.278
	100	0.594	1.959	5.593	0.596	1.960	5.619

We also investigated the percentage coverage of nominal 95% credible intervals. The results are presented in Table 4. As far as this simulation study is concerned, the intervals for the parameters of interest have in general the correct coverage percentages for both models investigated and these results do not depend on whether we have 60 or 100 areas. However, with a small number of areas we could face convergence problems unless we tighten the priors for the hyperparameters. The simulation study reveals that the population prediction is not affected by the forms of the vague priors assigned to the variance of the intercept term.

Table 4
The coverage rates of nominal 95% credible intervals for the population densities

Model	Num. of Areas	Gamma prior coverage(%)	Uniform prior coverage(%)
Spatial	60	96	96
	100	96	96
Hierarchical	60	94	94
	100	95	95

We analyzed the model fit when data generated from a model were fitted by the correct and the wrong models. Figure 2 presents the mean square error for the following situations: (a) data generated from the Spatial model and fitted by the Spatial and Hierarchical models and (b) data generated from the Hierarchical model and fitted by the Spatial and Hierarchical models. Since the form of the priors assigned to the parameters related to the variance does not affect the inference, we set uniform priors for both models. The ARE measures are shown in Figure 3.

It can be seen from Figure 2 that when the data are generated from the simpler model (Hierarchical) the more complex estimation procedures (Spatial) do not suffer any appreciable worsening of efficiency. On the other hand when the data are generated from the more complex model (Spatial) the simpler estimator (Hierarchical) has some inferior properties. However, this result does not hold for the ARE measurements. Figure 3 shows that fitting the model not used for generating the data results in appreciable increase in the relative bias. As it might be expected, model fitting and diagnostics are crucial in order to get suitable prediction of the small area population.

4.2 Application to real data

The PNAD data sets from 1992 to 1999, (excluding 1994 and 1996) and the population census data of 1991, 1996 and 2001 were used in our application. Our areas of interest are all the municipalities in São Paulo State, a total of 572 areas, of which 111 areas were sampled by the PNAD survey. Figure 4 shows the areas sampled by the PNAD, classified by the sampling definition: areas belong to metropolitan regions and self-representing areas (sampled with probability equal to 1) and non-self-representing areas. It should be noted that the census and PNAD have different periods of reference. We set $t = 0$ for the 1991 census. Thus, the values of t for the data provided by the PNAD are equal to the number of years between the reference period of the 1991 census and the respective PNAD. For instance, a survey datum provided by the PNAD 18 months after the 1991 census corresponds to $t = 1.5$.

Figure 5 shows the estimated coefficient of variation of the direct estimator by areas' sample sizes. These estimates are based on PNAD data. It can be seen that these coefficients of variation vary considerably with the areas and tend to decrease as the sample size increases. The high values of these coefficients show the difficulty in using only the direct estimator to provide municipal estimates. Furthermore, we cannot make any prediction for nonsampled areas by using only the direct estimators.

4.3 Specification of the prior distributions

The mean of the normal prior distributions of the parameters α , β and γ , related to the population evolution, were assigned by first expanding the function $\alpha + \beta \exp(\gamma t)$ around zero in a Taylor series up to the second order and then equating the resulting expression to the values of the mean density in the 1991 and 2000 censuses and the 1996 population count. In the absence of prior information, we considered a reasonably large value (10^6) for the prior variances of α , β and γ . Thus, we set $\alpha \sim U(-\infty, +\infty)$ (see Section 3.3 for further details), for the Spatial Model and $\alpha \sim N(370, 10^6)$, for the Hierarchical model and $\beta \sim N(726, 10^6)$, $\gamma \sim N(0.04, 10^6)$ for both models. The reason for this adjustment is to obtain a reasonable value of the prior means, but one that is essentially vague. Regarding the precisions and η_0, η_1 , we assigned relatively vague priors: $\tau_\alpha^2 \sim \text{Ga}(0.001, 0.001)$, $\tau_\gamma^2 \sim \text{Ga}(0.001, 0.001)$, $\eta_0 \sim N(0, 10^6)$ and $\eta_1 \sim N(0, 10^6)$.

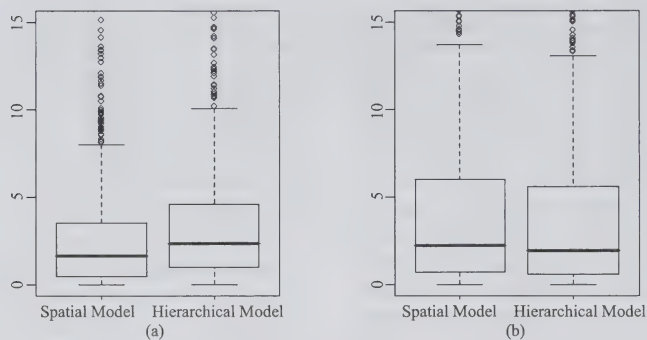


Figure 2 Box plots of mean square error (MSE) for the cases: (a) data generated from the Spatial model and respectively fitted by the Spatial and Hierarchical models and (b) data generated from the Hierarchical model and respectively fitted by the Spatial and Hierarchical models

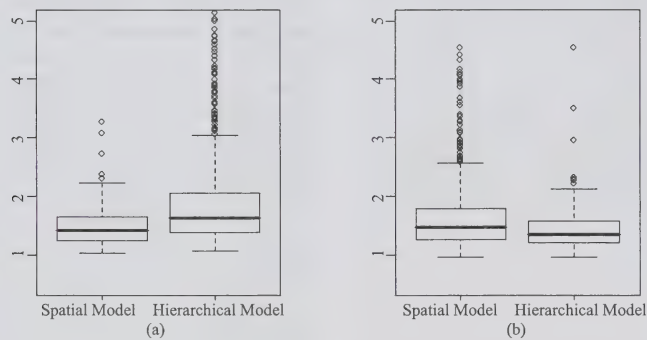


Figure 3 Box plots of absolute relative error (ARE) for the cases: (a) data generated from the Spatial model and respectively fitted by the Spatial and Hierarchical models and (b) data generated from the Hierarchical model and respectively fitted by the Spatial and Hierarchical models

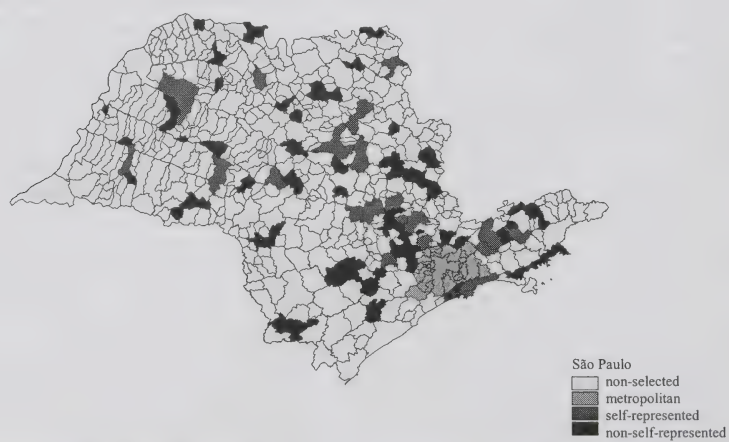


Figure 4 São Paulo municipalities sampled by the PNAD classified by the sampling definition

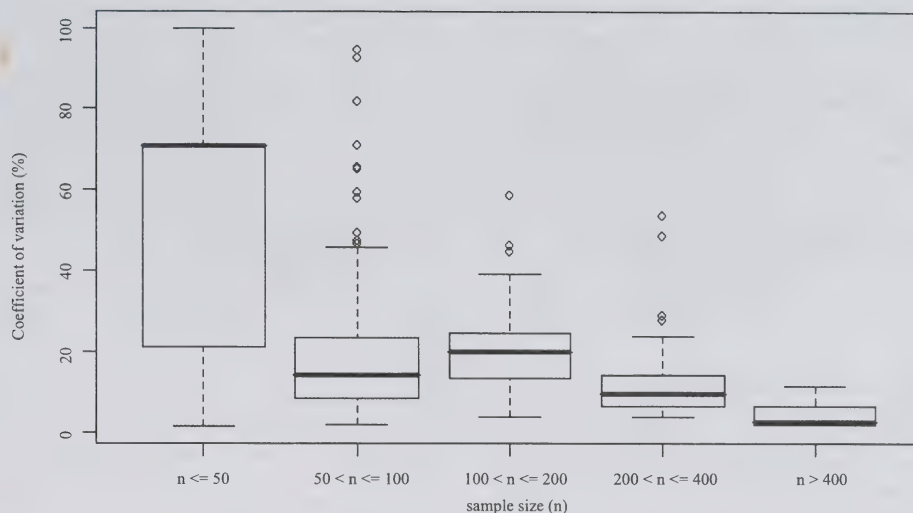


Figure 5 Boxplot of the coefficients of variation of the direct population estimates

4.4 Some results

We generated 20,000 samples after discarding the first 5,000. There is no evidence for non-convergence of the Hierarchical and the Spatial model parameters. A careful analysis of the MCMC outputs suggests that convergence was achieved for all model parameters. We summarize the results obtained by fitting the Hierarchical model (3) to the data provided by the PNAD survey. The posterior means of the model parameters were used as the point estimates. Table 5 presents these estimates together with the respective square root of the posterior variance. It can be seen from Table 5 that the estimate of η_1 is significantly positive, which agrees with what is expected by equation 4: the greater the sample size, the smaller σ_{it}^2 .

Table 5
Summary of the model (2) parameter posterior distributions

parameter	posterior mean	posterior std
α	892.500	202.000
β	105.700	1.278
γ	0.072	0.008
η_0	10.620	0.133
η_1	3.185	0.484
τ_α^2	2.174E-7	2.961E-8
τ_γ^2	139.000	19.560

Figure 6 shows that the posterior means of the parameters α and γ that index the hierarchical model seem to be spatially distributed. The parameters of neighboring areas seem more alike than those of distant areas, which suggests applying the Spatial model.

4.5 Model selection

The Expected Prediction Deviance (EPD) (Gelfand and Ghosh 1998) measure was applied to help choose the most suitable model. The EPD measure is the sum of two terms. The first term, denoted by G , can be interpreted as a goodness-of-fit measure and the second term, denoted by P , as a penalty term for underfitted as well as overfitted models. The respective expressions for G and P are given by: $G = \sum_{i=1}^m \sum_{t=1}^n (y_{it} - E(y_{it}^{rep}|M))^2$ and $P = \sum_{i=1}^m \sum_{t=1}^n V(y_{it}^{rep}|M)$, where the expectations and the variances are with respect to the posterior predictive distribution associated with a future observation (y_{it}^{rep}) of y_{it} generated under the assumed model (M). According to this criterion, the smaller its value, the better the model. As can be seen in Table 6, the EPD criterion slightly favors the Spatial model.

4.6 Analysis of the results

The most disaggregated level for which the PNAD provides precise estimates is the metropolitan region, which is a set of contiguous municipalities. In order to validate the results obtained with the spatial model, population estimates for the greater São Paulo metropolitan region were

compared to the official statistics projections. The posterior distribution of $\mu_r = \sum_{i=1}^r \pi_{it} * A_i$ is easily obtained by adding $\mu_r^{(t)} = \sum_{i=1}^r \pi_{it}^{(t)} * A_i$ to the MCMC algorithm, where μ_r represents the total population of the metropolitan region at time t and r is the number of municipalities belonging to that metropolitan region.

Table 6
Measures for selecting models for demographic density

Model	G	P	EPD
Hierarchical	1.37E+09	6.14E+09	7.51E+09
Spatial	1.05E+09	6.19E+09	7.24E+09

Figure 7 compares the population estimates (μ_r) of the São Paulo metropolitan region obtained by the Spatial model and the official statistics. The solid lines represent the limits of the 95% credible intervals of μ_r , while the dotted line shows the respective point estimates. The symbol (+) represents the observed official statistics. It is noteworthy that some official statistics projections are outside of the credibles inferior limit (including the 1991 Census). This indicates that further investigations should be made in order to find out the reasons for these discrepancies. However, when we compare them at municipality level, the overall conclusion is that the model predictions and official statistics reasonable agree. The 95% credible intervals contains 92.4% of the official statistics projections. The average of the absolute relative error (ARE) between the estimated population density and the official statistics projection are 3%. These ARE measures are on average nearly the same for selected and non-selected municipalities.

Figure 8 compares the point estimates of the population sizes (μ_{it}) with the official projection statistics and the official census population sizes for a sampled municipality. The official projection methodology assumes that a set of small areas and a larger area, which contains them, have the same population growth rate pattern. The population of the larger area is projected by a component method and then proportionally allocated to the small areas. The component method uses data from the most recent census as well as the number of births and deaths and net migrations obtained from administrative records. The component method projects the population for a time t by adding the population in a previous time with the number of births and

net migrations and subtracting the number of deaths in the same time interval.

The solid lines represent the 95% credible intervals for μ_{it} obtained by the Spatial model, while the dotted line shows the respective posterior means. The symbol (+) represents the official population projection for the intercensus period and the observed population in the census years. It is noteworthy that the point estimates are relatively close to the official projection statistics and the population obtained in the census year. This indicates that the use of the proposed model yields reliable estimates at municipality levels, with the extra advantage of providing a measure of the respective error.

We also analyze the estimates obtained for some municipalities not sampled in the PNAD. Figure 9 shows the model predictions, the 95% credible intervals, the official projection statistics and the observed population values in the censuses for a non-sampled municipality (+). It can be seen that the predictions obtained by the Spatial model reasonably agree with the official figures.

5. Final remarks

The model used in this article identifies the population growth trend of the municipalities. Reasonable estimates of the municipal populations are obtained for years with survey data, as well as for the years where census data are available. The point estimates have good precision and reasonably agree with estimates obtained for larger areas using other technique. The past information can be updated as soon as estimates become available from a new census or survey. Furthermore, the proposed approach provides the probability distribution of the quantity of interest, aiding the decision-making process.

Further work should be done in order to allow for autocorrelation of the parameters of interest over time. Extra information about the sampling variance estimates of the direct estimators could also be regarded as additional data. The assumption that the census coverage error is distributed symmetrically around zero could be relaxed by assigning a non-symmetric distribution to it. However a good knowledge of the shape of the distribution is required, which might be difficult in practice.

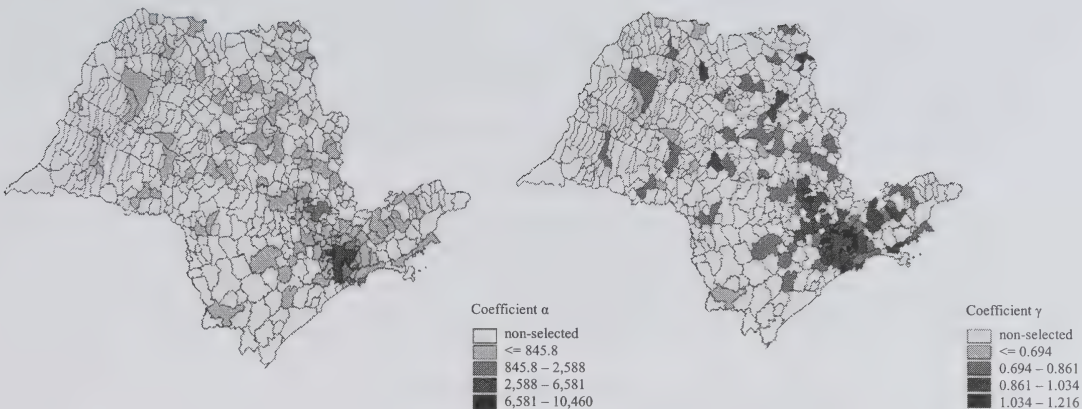


Figure 6 Posterior means of the parameters α and γ obtained by the hierarchical model

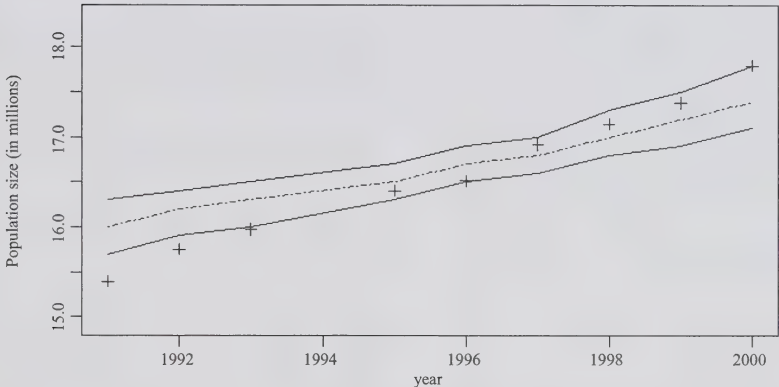


Figure 7 Comparison between the population sizes predicted by the spatial model and the official statistics (+) for the metropolitan region

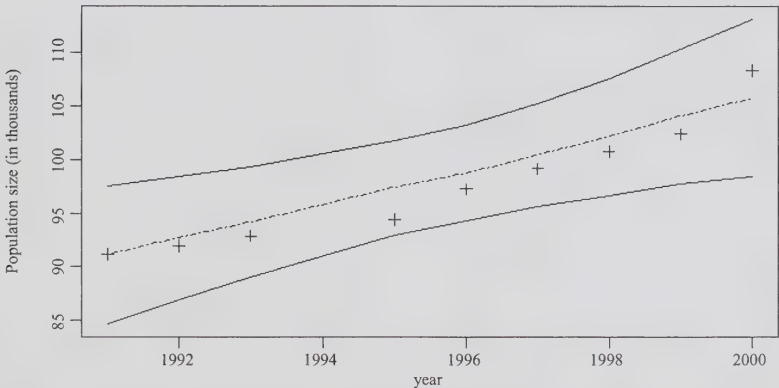


Figure 8 Comparison between the population sizes predicted by the spatial model and the official statistics (+) for a sampled municipality

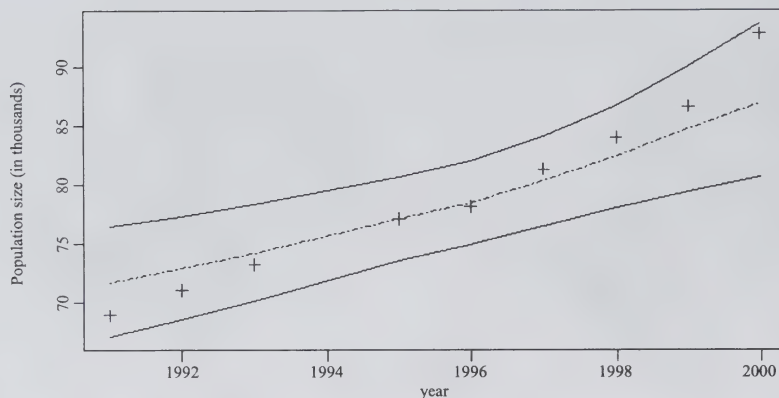


Figure 9 Population sizes predicted by the spatial model and the official statistics (+) for a non-sampled municipality

Acknowledgements

The authors thank an associate editor and two reviewers for their constructive comments and suggestions. The work of Fernando Moura and Helio Migon was funded in part by a research grant from the Brazilian National Council for the Development of Science and Technology (CNPq).

References

- Besag, J., and Kooperang, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.
- Brooks, S.P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 4, 434-455.
- Gelfand, A.E., and Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85, 1, 1-11.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 3, 515-533.
- Migon, H., and Gamerman, D. (1993). Generalized exponential growth models: A bayesian approach. *Journal of Forecasting*, 12, 573-584.
- Mollié, A. (1996). Bayesian mapping of disease. In: Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. Markov Chain Monte Carlo in Practice. New York: Chapman & Hall, 359-379.
- Moura, F.A.S., and Migon, H.S. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modeling: An International Journal*, 2, 3, 183-201.
- Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review*, 70, 1, 125-143.
- Pfeffermann, D., Moura, F.A.S. and Silva, P.L.N. (2006). Multi-level modeling under informative sampling. *Biometrika*, 93, 4, 943-959.
- Pfeffermann, D., and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of American Statistical Association*, 102, 1427-1439.
- Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.
- Souza, D.F. (2004). *Estimação de População em Nível Municipal via Modelos Hierárquicos e Espaciais*. Unpublished master's dissertation. Universidade Federal do Rio de Janeiro.
- Spiegelhalter, D.J., Thomas, A., Best, N. and Lunn, D. (2004). WinBUGS User Manual Version 1.4. MRC Biostatistics Unit, Cambridge.

Variance estimation in the presence of nonrespondents and certainty strata

Jun Shao and Katherine J. Thompson¹

Abstract

Business surveys often use a one-stage stratified simple random sampling without replacement design with some certainty strata. Although weight adjustment is typically applied for unit nonresponse, the variability due to nonresponse may be omitted in practice when estimating variances. This is problematic especially when there are certainty strata. We derive some variance estimators that are consistent when the number of sampled units in each weighting cell is large, using the jackknife, linearization, and modified jackknife methods. The derived variance estimators are first applied to empirical data from the Annual Capital Expenditures Survey conducted by the U.S. Census Bureau and are then examined in a simulation study.

Key Words: Covariate dependent nonresponse; Jackknife; Linearization; Ratio adjustment; Uniform nonresponse.

1. Introduction

Many business surveys use a one-stage stratified simple random sample without replacement design. Because of the skewness of the sampled populations, these designs generally include both certainty and non-certainty strata. With such designs, the sampling rates in the non-certainty strata are generally negligible (e.g., less than 20 percent in all strata). However, if the ultimate sampling unit is large business entity such as a company, the size of the universe is much smaller and often sampling fractions should not be ignored in computation of variance estimates.

Most surveys have nonresponse. We consider surveys using weighting adjustment for nonresponse. For certainty strata, there is no sampling error and, hence, standard variance formulas do not include any component for certainty strata. When nonresponse is present, however, there is an estimation error even in a certainty stratum, which is often an appreciable component of the total estimation error.

The purpose of this paper is to develop some methods for variance estimation that take into account the weighting adjustment for nonresponse and the existence of certainty strata. After introducing notation and assumptions in Section 2, we show that the jackknife and linearization variance estimators ignoring nonresponse in certainty strata, which are often currently used in many surveys, underestimate the true variance of the weight adjusted estimated population total. By directly deriving an approximate variance formula, we obtain two consistent variance estimators. These variance estimators are also consistent if there are non-certainty strata with large sampling fractions. A modified jackknife variance estimator taking into account the variability due to nonresponse in certainty strata is also derived.

In Section 3, we compare variance estimators using five years' of data from the Annual Capital Expenditures Survey (ACES) conducted by the U.S. Census Bureau. Simulation results are presented in Section 4 using a population generated from 2003 ACES data. Our simulation results show that the variance estimators ignoring certainty strata have large negative biases; the derived consistent variance estimators perform well when stratum sample sizes are all large and perform inconsistently otherwise; and the jackknife variance estimator ignoring all sampling fractions overestimates. Some concluding remarks are given in Section 5.

2. Main results

Consider a stratified sample without replacement from a finite population containing H strata. Let n_h and N_h be the sample and population size of stratum h , respectively, y_{hj} be a variable of interest that may have nonresponse, and x_{hj} be a covariate that takes positive values and does not have nonresponse, where j is the index of population unit and h is the index for stratum. Using the sample-response path considered by Fay (1991) and Shao and Steel (1999), we view the finite population as a census with y, x values and nonrespondents, i.e., each unit j in stratum h of the finite population is associated with an indicator I_{hj} ($= 1$ if y_{hj} is a respondent and $= 0$ if y_{hj} is a nonrespondent). Our sample is taken from this finite population, and if unit j in stratum h is in the sample, y_{hj} is a respondent if $I_{hj} = 1$ and a nonrespondent if $I_{hj} = 0$.

Let E_s and V_s be the expectation and variance, respectively, with respect to sampling and E_m, V_m , and P_m be the expectation, variance, and probability, respectively,

1. Jun Shao, University of Wisconsin-Madison and U.S. Census Bureau; Katherine J. Thompson, U.S. Census Bureau. E-mail: shao@stat.wisc.edu.

with respect to the model m specified in one of the following assumptions.

Assumption M. Values of (y_{hj}, x_{hj}, I_{hj}) in the finite population are independently generated from a superpopulation model m . The finite population is divided into P sub-populations such that, within sub-population p , the response probability $P_m(I_{hj} = 1 | y_{hj}, x_{hj}) = P_m(I_{hj} = 1 | x_{hj}) > 0$, $E_m(y_{hj} | x_{hj}) = \beta_p x_{hj}$, and $V_m(y_{hj} | x_{hj}) = \sigma_p^2 x_{hj}$, where β_p and σ_p are unknown parameters depending on p .

Assumption P. The finite population is divided into P sub-populations such that, under a superpopulation model, $P_m(I_{hj} = 1 | y_{hj}, x_{hj}) = \pi_p > 0$ is constant within sub-population p .

The sub-population in Assumption M or Assumption P is called nonresponse adjustment weighting cell (or weighting cell for short), since we handle nonrespondents by weight adjustment within each weighting cell. (If imputation is applied within each sub-population, then sub-populations are called imputation cells.) In applications, weighting cells may be strata, or unions of strata (strata are collapsed when they have insufficient respondents), or may cut across strata. Assumption M involves a prediction model between y_{hj} and x_{hj} and a covariate-dependent response mechanism within each weighting cell. The response mechanism under Assumption P is the within-weighting-cell uniform response mechanism and is often referred to as the quasi-random response model. Assumption P is stronger than Assumption M in terms of the response mechanism. However, Assumption M requires an explicit model between y_{hj} and x_{hj} within each weighting cell. In this paper we assume either Assumption M or Assumption P. Estimators that can be justified under Assumption P are referred to as the “quasi-randomization” estimators (Oh and Scheuren 1983).

When we study asymptotic consistency of estimators, we consider the limiting process of $k_p \rightarrow \infty$ for all p with fixed H and P , where k_p is the sample size in weighting cell p . If weighting cells are the same as strata or unions of strata, then $k_p \rightarrow \infty$ is the same as $n_h \rightarrow \infty$ for all h .

After the ratio-adjustment for nonresponse, we consider the following estimator of the total of y -values in the finite population:

$$\hat{Y} = \sum_p \sum_h \sum_{j \in s_h} \left(\frac{\hat{X}_p}{\hat{X}_{pr}} w_{hj} \right) \delta_{phj} I_{hj} y_{hj} = \sum_p \frac{\hat{X}_p}{\hat{X}_{pr}} \hat{Y}_{pr}, \quad (1)$$

where p is the index for weighting cell, s_h is the sample in stratum h , δ_{phj} is the indicator for the weighting cell p , and w_{hj} is the survey weight constructed for the stratified sampling,

$$\hat{X}_p = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} x_{hj}, \quad \hat{X}_{pr} = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} I_{hj} x_{hj},$$

and

$$\hat{Y}_{pr} = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} I_{hj} y_{hj}.$$

In the special case where weighting cells are the same as strata,

$$\hat{Y} = \sum_h \frac{\hat{X}_h}{\hat{X}_{hr}} \hat{Y}_{hr}, \quad (2)$$

where

$$\hat{X}_h = \sum_{j \in s_h} w_{hj} x_{hj}, \quad \hat{X}_{hr} = \sum_{j \in s_h} w_{hj} x_{hj} I_{hj},$$

and

$$\hat{Y}_{hr} = \sum_{j \in s_h} w_{hj} y_{hj} I_{hj}.$$

When the covariate $x_{hj} \equiv 1$, \hat{Y} is referred to as the count estimator. The count estimator controls respondent estimates to frame population totals. When the weighting cells are the same as strata, the count estimator uses the unweighted cell response rates, as recommended in Vartivarian and Little (2002).

Under Assumption M or P,

$$E_m E_s (\hat{Y} - Y) = E_s E_m (\hat{Y} - Y) = 0,$$

where Y is the finite population total of y values, and the total variance

$$V_{m_s}(\hat{Y} - Y) = E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y].$$

Let $V_1 = E_m[V_s(\hat{Y})]$ and $V_2 = V_m[E_s(\hat{Y}) - Y]$. To estimate V_1 , it suffices to estimate the sampling variance $V_s(\hat{Y})$. Since \hat{Y} defined by (1) is a sum of ratios and each of \hat{X}_p , \hat{X}_{pr} , and \hat{Y}_{pr} is a weighted total of variables and indicators, we can apply the stratified jackknife variance estimator

$$v_{J1} = \sum_h \left(1 - \frac{n_h}{N_h} \right) \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\hat{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \hat{Y}_{(hk)} \right)^2 \quad (3)$$

(see Wolter 1985 or Shao and Tu 1995), where $\hat{Y}_{(hj)}$ is the jackknife analog of \hat{Y} when unit j in stratum h is deleted. Note that sampling fractions are incorporated in this formula. When $k_p \rightarrow \infty$ for all weighting cells, the standard result for the complete data case (see, e.g., Krewski and Rao 1981) implies that the jackknife estimator v_{J1} is consistent for the sampling variance $V_s(\hat{Y})$, under Assumption M or P. Since V_1 is the expectation of $V_s(\hat{Y})$, v_{J1} is also consistent for V_1 under some minor conditions.

Since the function in (1) is the sum of ratios and data in different weighting cells are independent, a linearization estimator of $V_s(\hat{Y})$ can be derived using Taylor's expansion.

When weighting cells are the same as strata, for example, \hat{Y} is given by (2) and is a separate ratio estimator whose linearization variance estimator can be obtained using standard techniques. An alternative way to derive a linearization variance estimator is to linearize the jackknife estimator v_{J1} (Thompson and Yung 2006). The resulting estimator is

$$v_{L1} = \sum_h \frac{n_h}{n_h - 1} \sum_{j \in s_h} \left\{ \sum_p \left[\frac{\hat{X}_p}{\hat{X}_{pr}} (\bar{e}_{ph} - w_{hj} e_{phj} I_{hj} \delta_{phj}) + \frac{\hat{Y}_{pr}}{\hat{X}_{pr}} (\bar{x}_{ph} - w_{hj} x_{hj} \delta_{phj}) \right]^2 \right\}, \quad (4)$$

where $e_{phj} = y_{hj} - (\hat{Y}_{pr}/\hat{X}_{pr})x_{hj}$, $\bar{e}_{ph} = n_h^{-1} \sum_{j \in s_h} w_{hj} e_{phj} I_{hj} \delta_{phj}$, and $\bar{x}_{ph} = n_h^{-1} \sum_{j \in s_h} w_{hj} x_{hj} \delta_{phj}$. The estimator in (4) is exactly the same as the standard linearization variance estimator for the separate ratio estimator in (2) when weighting cells are the same as strata. Like v_{J1} , v_{L1} is consistent for V_1 when $k_p \rightarrow \infty$ under Assumption M or P, which follows from the standard result for the complete data case (Krewski and Rao 1981).

Since ratio is a smooth function, under Assumption M or P,

$$E_s(\hat{Y}) = \sum_p E_s \left(\frac{\hat{X}_p \hat{Y}_{pr}}{\hat{X}_{pr}} \right) \approx \sum_p \frac{E_s(\hat{X}_p) E_s(\hat{Y}_{pr})}{E_s(\hat{X}_{pr})} = \sum_p \frac{X_p Y_{pr}}{X_{pr}},$$

where

$$\begin{aligned} X_p &= \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} x_{hj}, \\ X_{pr} &= \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} x_{hj}, \\ Y_{pr} &= \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} y_{hj}, \end{aligned}$$

and \mathcal{P}_h is the finite population in stratum h . Let Y_p be the same as X_p with x_{hj} replaced by y_{hj} . Then

$$V_2 = V_m[E_s(\hat{Y}) - Y] \approx \sum_p V_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right).$$

Note that V_2 is small if the nonresponse rate is low ($V_2 = 0$ if there is no nonresponse) or if the model under Assumption M is highly predictive. If the overall sampling fraction, $\sum_h n_h / \sum_h N_h$, converges to 0, then V_2/V_1 converges to 0 and, hence v_{L1} and v_{J1} are consistent estimators of the total variance $V_{m,s}(\hat{Y}) = V_1 + V_2 \approx V_1$. Note that V_1 does not contain the variation from certainty strata due to nonresponse. Because the y -values from certainty strata are influential in the total Y in many surveys, and because in applications it is difficult to tell how small $\sum_h n_h / \sum_h N_h$ has to be for the convergence $V_2/V_1 \rightarrow 0$ to take place, it is necessary to estimate V_2 .

Under Assumption M, let \tilde{E}_m , \tilde{V}_m , and \tilde{C}_m be the conditional expectation, variance, and covariance, respectively, given all x -values and response indicators. Since

$$\tilde{E}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) = 0,$$

we obtain

$$\begin{aligned} V_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) &= E_m \left[\tilde{V}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] + V_m \left[\tilde{E}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] \\ &= E_m \left[\tilde{V}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] \\ &= E_m \left[\frac{X_p^2}{X_{pr}^2} \tilde{V}_m(Y_{pr}) - 2 \frac{X_p}{X_{pr}} \tilde{C}_m(Y_{pr}, Y_p) + \tilde{V}_m(Y_p) \right] \\ &= E_m \left[\frac{X_p^2}{X_{pr}^2} \sigma_p^2 X_{pr} - 2 \frac{X_p}{X_{pr}} \sigma_p^2 X_{pr} + \sigma_p^2 X_p \right] \\ &= \sigma_p^2 E_m \left(\frac{X_p^2}{X_{pr}} - X_p \right). \end{aligned}$$

Under Assumption P, let V_m^I be the variance with respect to I_{hj} 's. Since

$$E_m^I \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \approx 0,$$

we obtain

$$\begin{aligned} V_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) &\approx E_m \left[V_m^I \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] \\ &\approx E_m \left[\frac{1 - \pi_p}{\pi_p} \sum_h \sum_{j \in \mathcal{P}_h} \delta_{hj} \left(y_{hj} - \frac{Y_p}{X_p} x_{hj} \right)^2 \right] \\ &\approx E_m \left[\left(\frac{X_p^2}{X_{pr}} - X_p \right) S_p^2 \right], \end{aligned}$$

where

$$S_p^2 = \frac{1}{X_{pr}} \sum_h \sum_{j \in \mathcal{P}_h} \delta_{hj} \left(y_{hj} - \frac{Y_p}{X_p} x_{hj} \right)^2.$$

Since X_p and X_{pr} can be estimated by \hat{X}_p and \hat{X}_{pr} , respectively, to estimate V_2 we only need to find an estimator of σ_p^2 or S_p^2 . Under Assumption M, a regression estimator of β_p is $\hat{Y}_{pr}/\hat{X}_{pr}$ and a consistent estimator of σ_p^2 based on regression residuals is

$$\hat{\sigma}_p^2 = \frac{1}{\hat{X}_{pr}} \sum_h \sum_{j \in s_h} \delta_{phj} I_{hj} w_{hj} \left(y_{hj} - \frac{\hat{Y}_{pr}}{\hat{X}_{pr}} x_{hj} \right)^2.$$

From the theory of sampling, $\hat{\sigma}_p^2$ is also a consistent estimator of S_p^2 under Assumption P. Hence, under Assumption M or P, a consistent estimator of V_2 is

$$v_{L2} = \sum_p \hat{\sigma}_p^2 \left(\frac{\hat{X}_p^2}{\hat{X}_{pr}} - \hat{X}_p \right). \quad (5)$$

The subscript L indicates that this estimator is based on linearization.

In some applications $\sum_h n_h / \sum_h N_h$ is negligible and non-response in noncertainty strata has negligible contribution to the variance component V_2 , i.e.,

$$V_2 \approx V_m \left[\sum_p \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} - Y_{cp} \right) \right], \quad (6)$$

where the subscript c stands for certainty strata,

$$\begin{aligned} X_{cp} &= \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{J}_h} \delta_{phj} x_{hj}, & X_{cpr} &= \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{J}_h} \delta_{phj} I_{hj} x_{hj}, \\ Y_{cp} &= \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{J}_h} \delta_{phj} y_{hj}, & Y_{cpr} &= \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{J}_h} \delta_{phj} I_{hj} y_{hj}, \end{aligned}$$

and \mathcal{C} is the collection of indices of certainty strata. A consistent jackknife estimator of V_2 can be obtained as follows. Note that X_{cp} , X_{cpr} , and Y_{cpr} are estimators, since $\mathcal{P}_h = s_h$ for $h \in \mathcal{C}$, but Y_{cp} is not an estimator because of nonresponse. Thus, we cannot apply the jackknife to the function $X_{cp} Y_{cpr} / X_{cpr} - Y_{cp}$. From the previous derivation we note that, under Assumption M,

$$\begin{aligned} V_2 &\approx E_m \tilde{V}_m \left[\sum_p \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} - Y_{cp} \right) \right] \\ &= E_m \left[\sum_p \left(1 - \frac{X_{cpr}}{X_{cp}} \right) \tilde{V}_m \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} \right) \right]. \end{aligned}$$

Similarly, under Assumption P, the result holds with \tilde{V}_m replaced by V_m^I . Hence, we can apply the jackknife to the estimator $X_{cp} Y_{cpr} / X_{cpr}$. Let

$$\tilde{Y} = \sum_p \sqrt{1 - \frac{X_{cpr}}{X_{cp}}} \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} \right)$$

and $\tilde{Y}_{(hj)}$ be the jackknife analog of \tilde{Y} after unit j in $h \in \mathcal{C}$ is deleted, when we treat $X_{cp} Y_{cpr} / X_{cpr}$ as estimators. Then a jackknife estimator of V_2 is

$$v_{J2} = \sum_{h \in \mathcal{C}} \frac{N_h - 1}{N_h} \sum_{j \in \mathcal{J}_h} \left(\tilde{Y}_{(hj)} - \frac{1}{N_h} \sum_{k \in \mathcal{P}_h} \tilde{Y}_{(hk)} \right)^2$$

($n_h = N_h$ and $s_h = \mathcal{P}_h$ when $h \in \mathcal{C}$). The factor $\sqrt{1 - X_{cpr}/X_{cp}}$ in the formula for \tilde{Y} makes the appropriate adjustment for nonresponse. Under Assumption P, $X_{cpr}/X_{cp} \approx \pi_p$ is the response rate, which can be view as a "sampling" fraction for certainty strata.

The resulting jackknife estimator of the total variance $V_1 + V_2$ is then $v_{J1} + v_{J2}$. Since $n_h = N_h$ (i.e., $1 - n_h/N_h = 0$) if stratum h is a certainty stratum, it is easy to see that $v_{J1} + v_{J2}$ is equal to

$$v_J = \sum_h \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\tilde{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \tilde{Y}_{(hk)} \right)^2, \quad (7)$$

where

$$\tilde{Y}_{(hj)} = \begin{cases} \tilde{Y}_{(hj)} & \text{if stratum } h \text{ is a} \\ & \text{certainty stratum} \\ \hat{Y}_{(hj)} \sqrt{1 - \frac{n_h}{N_h}} & \text{if stratum } h \text{ is not a} \\ & \text{certainty stratum.} \end{cases}$$

Compared with the jackknife variance estimator v_{J1} in (3), v_J in (7) addresses the variability due to nonresponse in certainty strata, whereas v_{J1} does not. Under (6) and Assumption M or P, v_J is consistent.

Finally, the jackknife estimator that ignores all sampling fractions is:

$$\tilde{v}_J = \sum_h \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\hat{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \hat{Y}_{(hk)} \right)^2. \quad (8)$$

This estimator seems to be conservative, although it is not theoretically justified.

In summary, we have the following estimators of the total variance $V_{m,s}(\tilde{Y})$:

1. The jackknife estimator v_{J1} defined in (3), which underestimates when V_2/V_1 is not negligible.
2. The linearization estimator v_{L1} defined in (4), which is asymptotically equivalent to v_{J1} .
3. $v_L = v_{L1} + v_{L2}$ with v_{L2} is defined in (5), which is consistent.
4. $v_{JL} = v_{J1} + v_{L2}$, which is asymptotically equivalent to v_L .
5. The jackknife variance estimator v_J defined in (7), which is consistent when (6) holds.
6. The jackknife estimator \tilde{v}_J .

Under stratified simple random sampling and Assumption P, v_L is approximately the same as the variance estimator obtained by treating the set of respondents as an additional phase of the stratified simple random sample (i.e., a two-phase sample design) and applying standard variance formula (when $x_{hj} \equiv 1$) or the variance formula for calibration estimators (Kott 1994, Särndal, Swensson and Wretman 1992, and Hidirolou and Särndal 1998). This variance estimator, however, is not consistent when Assumption P does not hold.

3. Empirical comparisons

In this section, we apply the variance estimators described in Section 2 to five years of empirical data from the employer component of the ACES introduced in Section 1. Section 3.1 provides background on the ACES analysis variables, sample design, and estimation procedures. Section 3.2 presents the empirical comparisons.

3.1 Background of ACES

The ACES collects data about the nature and level of capital expenditures in non-farm businesses operating within the United States. Respondents report capital expenditures, broken down by type (expenditures on Structures and expenditures on Equipment) for the calendar year in all subsidiaries and divisions for all operations within the United States.

The ACES universe contains two sub-populations: employer companies (ACE-1) and non-employer (ACE-2) companies. (A nonemployer company is one that has no paid employees, has annual business receipts of \$1,000 or more (\$1 or more in the construction industries), and is subject to federal income taxes. Most nonemployers are self-employed individuals operating very small unincorporated businesses, which may or may not be the owner's principal source of income). Different forms are mailed to sample units depending on whether they are ACE-1 companies or ACE-2 companies. New ACE-1 and ACE-2 samples are selected each year, both with stratified simple random sample without replacement designs. The ACE-1 sample comprises approximately seventy-five percent of the ACES sample (roughly 46,000 companies selected per year for ACE-1, and 15,000 selected per year for ACE-2). In the ACE-1 design, units are stratified into size-class strata within each industry on the sampling frame. There are five separate ACE-1 strata in each industry, consisting of one certainty stratum (referred to as stratum 10) and four non-certainty strata defined by company size within industry (denoted by 2A through 2D, ranked from largest to smallest within industry), with approximately 500 non-certainty strata in each year's design. Sampling fractions in the large-size class-within-industry strata (2A) can be fairly high: in most years, approximately 55% of the sample in 2A strata are sampled at rates between 0.5 and 1. Sampling fractions in the other three size class within-industry strata are usually less than 0.20. Design weights range from 1 to 1,000, depending on industry and size-class strata. The ACE-2 component is much less highly stratified, with between a total of six to eight size-class strata used each year, and sampling fractions less than 0.01 in all strata. Our empirical analysis is restricted to the ACE-1 component of the survey, which meets all of the conditions described in the previous section.

The ACES publishes total and year-to-year change estimates. Estimates are published for the entire survey, and by industry code as indicated by the respondent units (not necessarily the industry code on the sampling frame). If there is no nonresponse, variances are estimated using the delete-a-group jackknife variance estimator (Kott 2001). To account for unit nonresponse, the ACE-1 component uses the ratio-adjustment procedure presented in Section 2 with administrative payroll data as the auxiliary variable x . Weighting cells are the design strata, provided that there is at least one respondent in the cell. Cell collapsing is extremely rare and is hereafter ignored in this paper. More details concerning the ACES survey design, methodology, and data limitations are available on-line at <http://www.census.gov/csd/ace>.

Although the ACE-1 survey design is fairly typical for a business survey, the collected data are not. Smaller companies often report legitimate values of zero for capital expenditures, and consequently the majority of the estimates are often obtained from the certainty and large non-certainty (2A) companies. As the capital expenditures are further cross-classified, the incidence of reported zeros (especially among smaller companies) increases.

3.2 Comparisons

To assess the effect of the unit non-response weight adjustment procedure on the ACE-1 standard errors, we computed variance estimates from unit nonresponse adjusted ACE-1 data using the ratio estimator with payroll as the auxiliary variable, in four industries, each with high sampling rates in the large company non-certainty strata (2A). The selected industries represent a cross-section of the sectors represented in the ACES. These industries and their North American Industrial Classification System (NAICS) codes are: Oil and Gas Extraction (211100), Nonmetallic Mineral Mining and Quarrying (212300), Other Miscellaneous Manufacturing (339900), and Architectural, Engineering, and Related Services (541300). In subsequent tables and discussions, industries are referred to by their NAICS code.

Table 1 presents variance estimate comparisons using five years' of ACE-1 survey data for three characteristics: the total capital expenditures (Total), capital expenditures on structures (Structures), and capital expenditures on equipment (Equipment). For comparison, the variance estimates are presented as a ratio to v_{j1} in Table 1. The estimated totals are also included. (Note that these totals are not the same as the published estimates, since they are computed using the industry classification on the frame, not the industry classification provided by the respondent).

As expected, the jackknife estimator v_{j1} and the linearization jackknife estimator v_{L1} are very close for all

variables. The consistent variance estimators (v_L and v_{JL}) are all noticeably larger than their corresponding jackknife counterparts (v_{L1} and v_{J1}). In general, most capital expenditures are reported by certainty or large non-certainty companies, so effect on variance estimation of including non-respondent component in the variance estimator is noticeable. The jackknife estimator v_J , which adjusts for the effect of certainty strata, is generally between v_{J1} and v_{JL} . In some cases, v_J is equal to or very close to v_{J1} , indicating that the variability due to nonresponse mainly comes from non-certainty strata with large sampling fractions. The jackknife estimate \bar{v}_J , which ignores sampling fractions, is much larger than any other estimates.

4. Simulation results

In this section, we present a simulation study using data modeled from the ACE-1 industries presented in the previous section. Section 4.1 describes the simulation settings. Section 4.2 presents and summarizes the results.

4.1 Simulation settings

We modeled our population using respondent data from the 2003 data collection of the three key items collected by the survey (Total, Structures, and Equipment). Frame data for the auxiliary variable (payroll) were available for all units. The complete population data were generated using the SIMDAT algorithm (Thompson 2000) with modeling cells equal to sampling strata and population size equal to the original frame size in each cell. Table 2 provides sampling fractions and correlation coefficients with the payroll for the modeled data in each stratum.

In the simulation, stratified simple random samples were selected from the generated population. We examine the statistical properties of the six variance estimators described in Section 2 over repeated samples under the following two different response mechanisms applied to the sample data:

1. The covariate-dependent response mechanism obtained by randomly applying response propensities modeled from the survey data with payroll as the covariate, which yields very high probabilities of responding to the large units and very small probabilities to the small (non-certainty) units;
2. The within-stratum uniform response mechanism obtained by using the observed survey response rate as the within-stratum response probability.

On the average, response probabilities in the individual stratum within industry were 0.85, 0.76, 0.77, 0.76, and 0.68 for strata 10, 2A, 2B, 2C, and 2D, respectively.

We selected 5,000 samples from the population, computed \hat{Y} in (1) from each sample with nonresponse and weight adjustment, and computed the empirical mean and

variance of the 5,000 \hat{Y} values. This was done for each industry and each item, with two adjustment methods: the ratio and count estimators. When \hat{Y} is the ratio estimator using the payroll as the auxiliary variable, the absolute value of the empirical relative bias is under 1.4% and is smaller than 1% in most cases. For the count estimator under the within-stratum uniform response mechanism, its absolute value of the empirical relative bias is under 0.5%. The count estimator is not approximately unbiased in theory under the covariate-dependent response mechanism. In the simulation, however, its absolute value of the empirical relative bias is under 1% in most cases and has a maximum value of 2.7%. The empirical variance of the 5,000 \hat{Y} values was used as the “true value” of the variance of \hat{Y} .

4.2 Results

In 2,000 of the 5,000 samples, we computed the six different variance estimates for all three items, four industries, and two weight adjustment methods. We examined the statistical properties of each of variance estimator over repeated samples using the relative bias (RB) defined as

$$\frac{\text{the average of 2,000 variance estimates}}{\text{the true variance}} - 1,$$

the stability (ST) defined as

$$\frac{\sqrt{\text{the empirical mean squared error of variance estimate}}}{\text{the true variance}},$$

and the error rate (ER) defined as the empirical proportion of the approximate 90% confidence intervals ($\hat{Y} \pm 1.645\sqrt{\text{variance estimate}}$) from 2,000 samples that do not contain the true population total.

Tables 3 and 4 respectively report the simulation results under the two response mechanisms. The results from these tables can be summarized as follows.

1. Two variance estimators ignoring V_2 , v_{J1} and v_{L1} , have large negative relative biases in general. The error rates of the related confidence intervals are also large.
2. Two consistent variance estimators, v_L and v_{JL} , have very similar performances and are generally much better than v_{J1} and v_{L1} in terms of the relative bias and the error rate of the related confidence intervals.
3. The jackknife variance estimator v_J performs well in industries 339900 and 541300, but may have large positive relative biases in industries 211000 and 212300. We think that this is a “small sample” effect, since v_J is justified by asymptotic consistency and the sizes of the certainty strata in industries 211000 and 212300 are 26 and 30, respectively (Table 2). The sizes of the certainty

strata for the other two industries are 158 and 160, respectively. In fact, the performance of v_L and v_{JL} is generally better in industries 339900 and 541300.

4. In some cases v_J has more than 10% negative relative biases, which is caused by the fact that some non-certainty strata have large sampling fractions, *i.e.*, the approximation (6) does not hold enough.
5. The jackknife variance estimator \tilde{v}_J ignoring all sampling fractions has very large positive relative biases and is too conservative.

5. Concluding remarks

When nonresponse is present in certainty strata (or strata with large sampling fractions), the jackknife and the linearization variance estimators that ignore certainty strata (or strata with large sampling fractions) are not acceptable because of their large negative biases. We derive two asymptotically unbiased and consistent variance estimators

by adding an extra term that accounts the variability from nonresponse in certainty strata (or strata with large sampling fractions). We also derive a modified jackknife estimator that is consistent when the certainty strata are the only strata that contribute to the variance due to nonresponse (*i.e.*, Assumption (6) holds).

Our simulation results show that the three derived variance estimators perform well when stratum sample sizes are all large and perform inconsistently otherwise, and that the jackknife variance estimator that ignores all sampling fractions is very conservative.

Compared with the linearization method, the jackknife requires more computational resources but it has other advantages such as being easy to program, using a single recipe for different problems, and not requiring complicated or separate derivations for different estimators. Our linearization variance estimator given in (4) is in fact obtained by linearizing the jackknife estimator in (3).

Table 1
Variance estimates for \hat{Y} with ratio adjustment in ACE-1 survey

Industry	Item	Year	\hat{Y}	v_{J1}	$\frac{v_{L1}}{v_{J1}}$	$\frac{v_L}{v_{J1}}$	$\frac{v_{JL}}{v_{J1}}$	$\frac{v_J}{v_{J1}}$	$\frac{\tilde{v}_J}{v_{J1}}$
211000	Total	2002	1.63E+7	4.63E+11	0.97	1.14	1.17	1.00	17.3
		2003	2.28E+7	6.87E+12	0.95	1.21	1.26	1.00	2.81
		2004	2.30E+7	2.45E+12	0.98	1.23	1.25	1.00	4.77
		2005	3.08E+7	4.29E+12	0.98	1.22	1.24	1.19	4.77
		2006	4.18E+7	6.29E+12	0.99	1.17	1.19	1.00	8.78
	Structures	2002	1.31E+7	3.99E+11	0.97	1.14	1.17	1.00	15.3
		2003	1.86E+7	5.78E+12	0.94	1.22	1.27	1.00	2.78
		2004	1.70E+7	8.39E+11	0.99	1.42	1.43	1.00	11.3
		2005	2.64E+7	3.84E+12	0.98	1.22	1.24	1.16	4.64
		2006	3.55E+7	5.41E+12	0.99	1.19	1.21	1.00	8.76
	Equipment	2002	3.20E+6	6.14E+10	0.98	1.15	1.17	1.00	7.26
		2003	4.18E+6	8.39E+11	0.97	1.22	1.24	1.00	1.70
		2004	6.01E+6	1.54E+12	0.97	1.13	1.16	1.00	1.39
		2005	4.33E+6	1.34E+11	0.97	1.22	1.25	1.15	6.17
		2006	6.31E+6	7.14E+11	0.99	1.12	1.13	1.00	2.68
212300	Total	2002	1.56E+6	4.14E+10	0.81	1.06	1.24	1.20	3.19
		2003	1.33E+6	1.21E+10	0.94	1.18	1.24	1.36	5.43
		2004	2.01E+6	2.86E+10	0.96	1.60	1.65	2.20	6.04
		2005	1.96E+6	1.93E+10	0.98	1.12	1.14	2.30	6.04
		2006	2.28E+6	2.19E+10	0.96	1.26	1.30	3.22	11.7
	Structures	2002	2.22E+5	4.36E+8	1.00	1.11	1.11	1.64	8.61
		2003	1.49E+5	2.27E+8	0.96	1.28	1.32	1.48	7.32
		2004	4.14E+5	1.03E+8	0.96	46.6	46.6	75.3	426
		2005	2.23E+5	9.33E+8	0.99	1.12	1.13	1.32	1.88
		2006	2.20E+5	1.88E+9	0.97	1.20	1.22	1.19	2.29
	Equipment	2002	1.33E+6	4.05E+10	0.81	1.06	1.25	1.15	2.86
		2003	1.18E+6	1.13E+10	0.94	1.20	1.26	1.32	5.07
		2004	1.60E+6	2.82E+10	0.96	1.40	1.44	1.53	3.30
		2005	1.73E+6	1.62E+10	0.97	1.16	1.19	2.33	6.69
		2006	2.06E+6	2.14E+10	0.96	1.26	1.30	2.94	10.8

Table 1 (continued)
Variance estimates for \hat{Y} with ratio adjustment in ACE-1 survey

Industry	Item	Year	\hat{Y}	v_{j1}	$\frac{v_{j1}}{v_{j1}}$	$\frac{v_j}{v_{j1}}$	$\frac{v_{j2}}{v_{j1}}$	$\frac{v_j}{v_{j1}}$	$\frac{\bar{v}_j}{v_{j1}}$
339900	Total	2002	1.75E+6	1.94E+10	0.99	1.27	1.29	1.10	3.71
		2003	1.58E+6	2.99E+10	0.98	1.24	1.27	1.10	1.60
		2004	1.70E+6	1.00E+10	0.99	1.40	1.40	1.69	4.61
		2005	1.77E+6	2.55E+10	0.99	1.28	1.29	1.25	3.02
		2006	1.94E+6	5.51E+10	0.99	1.23	1.25	1.12	2.15
	Structures	2002	2.99E+5	1.21E+9	0.99	1.24	1.24	1.09	3.55
		2003	1.93E+5	8.54E+8	0.99	1.27	1.28	1.09	1.75
		2004	2.10E+5	2.00E+8	0.99	1.86	1.87	2.08	5.89
		2005	2.56E+5	5.07E+8	0.99	1.80	1.81	1.97	9.61
		2006	5.97E+5	4.93E+10	0.99	1.19	1.20	1.01	1.16
	Equipment	2002	1.45E+6	1.62E+10	0.99	1.27	1.28	1.07	3.02
		2003	1.39E+6	2.71E+10	0.97	1.24	1.27	1.09	1.58
		2004	1.49E+6	9.14E+9	0.99	1.40	1.41	1.62	4.61
		2005	1.51E+6	2.45E+10	0.99	1.22	1.23	1.15	2.12
		2006	1.34E+6	5.65E+9	0.99	1.42	1.43	1.60	6.20
541300	Total	2002	3.38E+6	2.32E+10	0.99	1.47	1.48	1.67	5.02
		2003	3.09E+6	2.61E+10	0.99	1.26	1.27	1.05	1.62
		2004	3.97E+6	1.12E+11	1.00	1.23	1.23	1.03	1.37
		2005	4.94E+6	2.54E+11	1.00	1.20	1.20	1.04	1.71
		2006	4.96E+6	2.82E+10	1.00	1.40	1.40	1.75	8.36
	Structures	2002	7.41E+5	6.32E+9	1.00	1.70	1.71	1.64	7.47
		2003	4.29E+5	3.32E+9	1.00	1.29	1.29	1.01	1.33
		2004	6.96E+5	4.38E+10	1.00	1.22	1.22	1.00	1.40
		2005	7.12E+5	9.00E+9	1.00	1.25	1.25	1.08	2.08
		2006	8.73E+5	3.44E+9	1.00	1.58	1.59	1.63	9.88
	Equipment	2002	2.96E+6	1.39E+10	0.99	1.37	1.38	1.54	3.95
		2003	2.66E+6	1.94E+10	0.99	1.25	1.26	1.05	1.59
		2004	3.27E+6	5.83E+10	1.00	1.22	1.23	1.04	1.29
		2005	4.23E+6	2.40E+11	1.00	1.19	1.20	1.03	1.59
		2006	4.09E+6	2.35E+10	1.00	1.27	1.28	1.49	5.47

Table 2
Population characteristics for the simulation study

Industry	Stratum	Population size	Sampling fraction	Correlation with Payroll		
				Total	Structures	Equipment
211000	10	26	1.00	0.65	0.53	0.95
	2A	128	0.77	0.68	0.66	0.22
	2B	372	0.11	0.57	0.51	0.51
	2C	1,800	0.02	-0.07	0.00	-0.10
	2D	10,406	0.00	0.28	0.00	0.28
212300	10	30	1.00	0.96	0.95	0.94
	2A	108	0.37	0.85	0.74	0.77
	2B	414	0.07	0.03	0.76	-0.03
	2C	1,310	0.03	0.42	0.13	0.43
	2D	4,762	0.01	0.44	-0.22	0.44
339900	10	158	1.00	0.80	0.40	0.80
	2A	498	0.26	0.40	0.04	0.51
	2B	2,048	0.05	0.20	0.24	0.18
	2C	6,310	0.02	0.19	0.48	0.09
	2D	25,288	0.00	0.37	0.67	0.36
541300	10	160	1.00	0.60	0.56	0.59
	2A	959	0.38	0.20	0.39	0.06
	2B	4,531	0.06	0.28	0.13	0.27
	2C	17,913	0.01	0.08	0.06	0.08
	2D	67,440	0.00	0.13	-0.01	0.15

Table 3
Simulation results (in %) for variance estimation under covariate-dependent response mechanism

Estimate	Item	Industry		v_{J1}	v_{L1}	v_L	v_{JL}	v_J	\bar{v}_J
Ratio	Total	211000	RB	-35.8	-38.1	-10.3	-8.0	39.1	113.9
			ST	49.8	50.4	47.4	48.6	252.9	182.9
			ER	19.6	19.8	12.2	11.8	10.7	1.1
		212300	RB	-20.4	-22.2	-4.48	-2.69	54.8	266.4
			ST	30.3	31.1	26.8	27.3	139.1	268.8
			ER	12.6	12.6	9.9	9.6	6.3	0.1
		339900	RB	-21.2	-22.5	0.26	1.55	-5.34	52.5
			ST	47.3	47.0	55.0	56.0	43.9	67.8
			ER	14.3	14.6	10.4	10.3	10.0	2.6
		541300	RB	-20.7	-21.0	3.83	4.08	-11.6	18.4
			ST	32.7	32.8	34.9	35.0	29.4	32.0
			ER	12.6	12.7	8.6	8.6	10.7	6.2
	Structures	211000	RB	-38.0	-40.1	-11.9	-9.59	33.9	108.1
			ST	51.3	51.9	48.5	49.6	244.4	180.8
			ER	20.9	21.1	12.9	12.6	11.1	1.1
		212300	RB	-23.2	-23.9	-12.4	-11.6	33.2	341.5
			ST	31.5	32.0	27.1	27.0	95.0	344.3
			ER	12.3	12.3	10.4	10.3	6.9	0.1
		339900	RB	-20.0	-20.4	-6.31	-5.88	-10.9	39.8
			ST	42.5	42.7	42.3	42.3	39.9	64.0
			ER	15.9	16.0	12.7	12.6	13.2	5.4
		541300	RB	-20.0	-20.1	0.09	0.33	-15.9	15.7
			ST	42.6	42.5	50.5	50.7	41.1	42.7
			ER	13.1	13.2	9.9	9.9	12.0	6.5
	Equipment	211000	RB	-15.0	-17.3	14.1	16.4	-9.37	27.9
			ST	63.9	62.6	87.7	90.0	64.1	69.6
			ER	16.2	16.7	13.3	13.0	14.7	6.7
		212300	RB	-21.4	-23.3	-4.13	-2.21	39.7	201.1
			ST	31.7	32.5	28.4	29.0	113.7	204.4
			ER	13.3	13.5	10.2	10.1	7.7	0.2
		339900	RB	-21.4	-22.8	1.18	2.57	-7.29	50.8
			ST	51.2	50.9	60.8	61.9	47.9	69.2
			ER	15.5	15.8	11.6	11.4	11.0	2.3
		541300	RB	-19.7	-19.9	6.16	6.43	-11.9	12.8
			ST	33.8	33.9	38.4	38.5	31.0	30.9
			ER	12.5	12.5	8.9	8.9	11.0	7.0
Count	Total	211000	RB	-30.1	-31.9	0.05	1.85	1.05	103.1
			ST	50.4	50.5	55.9	57.3	46.7	113.4
			ER	15.3	15.6	9.0	8.8	8.7	1.0
		212300	RB	-33.2	-34.6	-6.30	-4.96	17.6	204.5
			ST	38.7	39.6	27.7	27.8	42.8	208.6
			ER	14.1	14.7	9.1	8.7	6.9	0.4
		339900	RB	-23.9	-24.6	1.73	2.44	-14.2	46.4
			ST	47.5	47.4	55.2	55.7	43.4	62.4
			ER	13.4	13.5	9.1	9.1	10.7	2.5
		541300	RB	-22.9	-23.2	1.68	1.94	-18.8	15.4
			ST	32.9	33.0	32.0	32.2	30.2	28.9
			ER	11.5	11.6	7.2	7.1	10.6	5.2
	Structures	211000	RB	-30.3	-32.2	-0.15	1.65	-1.27	101.5
			ST	51.3	51.3	57.3	58.7	46.7	112.3
			ER	15.8	16.3	9.6	9.4	9.2	0.8
		212300	RB	-37.4	-38.0	-13.5	-12.9	3.53	250.2
			ST	41.6	42.0	28.9	28.8	32.2	254.8
			ER	15.4	15.6	9.6	9.5	8.1	0.4
		339900	RB	-20.0	-20.3	-4.33	-4.00	-14.5	38.6
			ST	42.3	42.4	42.4	42.4	40.1	62.8
			ER	14.6	14.7	11.9	11.8	13.6	5.0
		541300	RB	-20.9	-21.2	-0.54	-0.32	-18.9	14.5
			ST	41.6	41.6	47.8	48.0	40.6	40.9
			ER	12.5	12.5	9.1	9.1	12.1	6.0
	Equipment	211000	RB	-17.8	-20.0	11.2	13.3	-13.0	26.6
			ST	58.9	58.0	76.4	78.4	57.7	64.1
			ER	15.7	15.8	12.5	12.3	14.5	6.1
		212300	RB	-30.7	-32.2	-4.74	-3.27	12.1	164.3
			ST	37.6	38.6	29.1	29.3	38.7	168.9
			ER	14.1	14.5	9.6	9.5	7.9	0.6
		339900	RB	-24.1	-24.9	2.52	3.27	-15.2	45.0
			ST	51.2	51.1	61.0	61.5	47.7	64.1
			ER	14.8	15.1	9.9	9.8	11.9	2.3
		541300	RB	-21.6	-21.9	4.10	4.39	-18.1	10.1
			ST	33.6	33.7	35.2	35.3	31.5	28.2
			ER	11.1	11.1	7.2	7.1	10.3	5.9

Table 4
Simulation results (in %) for variance estimation under within-stratum uniform response mechanism

Estimate	Item	Industry		v_{J1}	v_{L1}	v_L	v_{JL}	v_J	\hat{v}_J
Ratio	Total	211000	RB	-49.2	-50.4	-17.2	-16.0	89.2	138.4
			ST	55.4	56.1	43.7	43.9	310.5	258.3
			ER	26.9	27.1	13.9	13.8	9.10	1.80
		212300	RB	-5.42	-7.99	16.1	18.7	111.7	337.2
			ST	28.9	28.5	37.4	39.6	179.2	341.7
			ER	13.5	13.8	9.85	9.50	5.85	0.05
		339900	RB	-9.37	-10.5	18.0	19.2	16.5	83.8
			ST	45.8	45.4	59.0	60.1	48.4	95.6
			ER	14.5	14.7	9.55	9.55	8.60	2.65
		541300	RB	-8.83	-9.03	18.2	18.4	6.62	44.5
			ST	26.7	26.8	36.6	36.7	28.2	52.3
			ER	12.6	12.6	8.45	8.45	9.70	5.35
	Structures	211000	RB	-52.6	-53.7	-19.2	-18.0	78.4	128.0
			ST	58.0	58.7	45.3	45.4	290.8	248.5
			ER	28.7	29.0	15.1	14.9	9.80	2.25
		212300	RB	-16.2	-18.1	9.92	11.9	63.5	356.2
			ST	32.0	32.4	37.1	38.5	108.6	361.9
			ER	15.5	16.0	10.9	10.7	6.65	0.35
		339900	RB	-13.2	-13.6	13.9	14.3	1.15	54.9
			ST	47.8	47.8	59.2	59.5	46.3	82.7
			ER	17.2	17.2	12.6	12.5	13.8	6.40
		541300	RB	-8.9	-9.2	19.2	19.5	-2.22	36.0
			ST	39.4	39.3	53.7	54.0	38.6	55.9
			ER	12.9	12.9	8.85	8.85	11.3	6.85
	Equipment	211000	RB	-1.1	-2.75	27.5	29.1	12.8	60.1
			ST	64.4	63.2	88.6	90.3	71.1	89.8
			ER	15.3	15.6	12.0	12.0	12.7	5.10
		212300	RB	-6.3	-8.96	16.8	19.4	90.0	263.1
			ST	30.3	29.8	39.3	41.6	148.6	269.1
			ER	13.9	14.2	10.1	9.60	6.75	0.15
		339900	RB	-8.84	-10.1	19.5	20.7	15.8	84.6
			ST	50.8	50.3	65.7	66.9	52.9	98.8
			ER	15.1	15.3	10.3	10.3	9.50	2.45
		541300	RB	-6.89	-7.1	19.9	20.1	6.76	38.5
			ST	28.6	28.6	40.0	40.1	30.1	48.3
			ER	12.4	12.4	8.60	8.55	10.3	5.90
Count	Total	211000	RB	-27.8	-29.0	14.2	15.4	16.3	149.5
			ST	47.4	47.5	53.1	54.2	44.4	158.1
			ER	16.0	16.2	8.30	8.30	7.45	1.85
		212300	RB	-33.5	-34.9	15.3	16.7	23.9	219.9
			ST	40.0	40.9	38.5	39.5	39.4	228.5
			ER	18.8	19.3	9.80	9.65	8.55	1.90
		339900	RB	-16.5	-17.1	20.2	20.8	4.21	75.7
			ST	45.1	44.0	57.9	58.5	42.4	87.6
			ER	15.6	15.8	9.40	9.35	10.8	3.20
		541300	RB	-9.61	-9.81	18.9	19.1	-0.77	45.0
			ST	26.5	26.5	36.0	36.1	24.7	52.2
			ER	12.4	12.4	8.55	8.55	11.3	4.85
	Structures	211000	RB	-27.5	-28.7	14.5	15.7	14.6	149.0
			ST	48.1	48.1	54.5	55.6	45.1	157.6
			ER	17.1	17.5	9.05	9.00	8.50	2.05
		212300	RB	-39.4	-40.4	11.6	12.6	10.1	238.5
			ST	44.8	45.5	38.8	39.6	32.1	248.4
			ER	20.2	20.7	9.95	9.85	10.3	1.80
		339900	RB	-14.2	-14.6	13.6	14.0	-3.55	53.5
			ST	47.1	47.0	57.3	57.6	45.1	80.5
			ER	17.6	17.7	12.1	12.1	14.7	6.30
		541300	RB	-9.54	-9.76	20.0	20.2	-5.32	36.0
			ST	39.1	39.0	53.3	53.5	38.3	55.8
			ER	12.6	12.6	9.05	9.05	11.9	6.55
	Equipment	211000	RB	-8.12	-9.64	22.7	24.2	1.56	54.3
			ST	58.0	57.1	76.2	77.7	57.5	82.1
			ER	16.5	16.7	12.4	12.4	14.6	6.45
		212300	RB	-28.5	-30.0	17.1	18.6	21.5	189.7
			ST	37.6	38.4	40.9	42.0	38.2	198.9
			ER	18.1	18.5	9.95	9.80	9.25	1.75
		339900	RB	-15.8	-16.4	21.8	22.5	4.69	76.8
			ST	49.5	49.3	64.6	65.2	47.4	91.1
			ER	16.4	16.5	9.45	9.40	11.4	3.20
		541300	RB	-7.53	-7.74	20.2	20.4	0.26	38.8
			ST	28.2	28.2	39.2	39.4	27.2	48.0
			ER	12.7	12.7	8.30	8.25	11.3	5.65

Acknowledgements

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical or methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors thank two referees and an associate editor for their helpful comments and suggestions, and Carol Caldwell, Rita Petroni, and Mark Sands for their useful comments on an earlier version of this paper. Jun Shao's research was partially supported by the NSF Grant SES-0705033.

References

- Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429-440.
- Hidirolou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.
- Kott, P. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.
- Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment of unit nonresponse. *Incomplete Data in Sample Surveys*. New York: Academic Press, 20, 143-184.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- Thompson, J.R. (2000). *Simulation: A Modeler's Approach*. New York: John Wiley & Sons, Inc.
- Thompson, K.J., and Yung, W. (2006). To replicate (a weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.
- Vartivarian, S., and Little, R.J. (2002). On the formation of weighting adjustment cells for unit non-response. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3553-3558.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Rescaled bootstrap for stratified multistage sampling

John Preston¹

Abstract

In large scaled sample surveys it is common practice to employ stratified multistage designs where units are selected using simple random sampling without replacement at each stage. Variance estimation for these types of designs can be quite cumbersome to implement, particularly for non-linear estimators. Various bootstrap methods for variance estimation have been proposed, but most of these are restricted to single-stage designs or two-stage cluster designs. An extension of the rescaled bootstrap method (Rao and Wu 1988) to stratified multistage designs is proposed which can easily be extended to any number of stages. The proposed method is suitable for a wide range of reweighting techniques, including the general class of calibration estimators. A Monte Carlo simulation study was conducted to examine the performance of the proposed multistage rescaled bootstrap variance estimator.

Key Words: Bootstrap; Calibration; Multistage sampling; Stratification; Variance estimation.

1. Introduction

Stratified multistage sampling designs are especially suited to large scaled sampled surveys because of the advantage of clustering collection effort. Various methods exist for variance estimation for these complex survey designs. The most commonly used methods are the linearization (or Taylor) method, and resampling methods, such as the jackknife, balance repeated replication and the bootstrap. The linearization method can be quite cumbersome to implement for complex survey designs as it requires the derivation of separate variance formulae for each non-linear estimator. Some approximations are normally required for the variance of non-linear functions, such as ratios and correlation and regression coefficients, and functionals, such as quantiles.

On the other hand, the various resampling methods employ a single variance formulae for all estimators. The replication methods can reflect the effects of a wide range of reweighting techniques, including calibration, and adjustments due to provider non-response and population under-coverage. The jackknife and balance repeated replication methods are only applicable to stratified multistage designs where the clusters are sampled with replacement or the first-stage sampling fractions are negligible. A number of different bootstrap methods for finite population sampling have been proposed in the literature, including the with-replacement bootstrap (McCarthy and Snowden 1985), the rescaled bootstrap (Rao and Wu 1988), the mirror match bootstrap (Sitter 1992a), and the without-replacement bootstrap (Gross 1980; Bickel and Freedman 1984; Sitter 1992b). A summary of these bootstrap methods can be found in Shao and Tu (1995).

Most of these bootstrap methods are restricted to single-stage designs or multistage designs where the first-stage sampling units are selected with replacement or the

first-stage sampling fractions are small in most strata. However, in many large scaled sample surveys it is common practice to employ highly stratified multistage designs where units are selected using simple random sampling without replacement at each stage. Some typical examples of these types of surveys are employer-employee surveys, such as the Survey of Employee Earnings and Hours (ABS 2008), and school-student surveys, such as the National Survey on the Use of Tobacco by Australian Secondary School Students (White and Hayman 2006).

McCarthy and Snowden (1985) proposed an extension of their with-replacement bootstrap to two-stage sampling in the special case of equal cluster sizes and equal within cluster sample sizes, while Rao and Wu (1988) and Sitter (1992a) have given extensions of their rescaled bootstrap and mirror match bootstrap methods to two-stage cluster sampling. More recently, Funaoka, Saigo, Sitter and Toida (2006) proposed two Bernoulli-type bootstrap methods, the general Bernoulli bootstrap and the short cut Bernoulli bootstrap, which can easily handle multistage stratified designs where units are selected using simple random sampling without replacement at each stage. The general Bernoulli bootstrap has the advantage that it can handle any combination of sample sizes, but it requires a much larger number of random number generations than the short cut Bernoulli bootstrap.

In this paper, an extension of the rescaled bootstrap procedure to stratified multistage sampling where units are selected using simple random sampling without replacement at each stage is proposed. In Section 2, the notation for stratified multistage sampling is introduced. In Section 3, the extension of the rescaled bootstrap estimator to multistage sampling is described. The main findings of a simulation study are reported in Section 4. Some concluding remarks are provided in Section 5.

1. John Preston, Australian Bureau of Statistics, 639 Wickham Street, Fortitude Valley QLD 4006, Australia. E-mail: john.preston@abs.gov.au.

2. Stratified multistage sampling

For simplicity, the case of stratified three-stage sampling is presented. Consider a finite population U divided into H nonoverlapping strata $U = \{U_1, \dots, U_H\}$, where U_h is comprised of N_{1h} primary sampling units (PSU's). At the first-stage, a simple random sample without replacement (SRSWOR) of n_{1h} PSU's are selected with selection probabilities $\pi_{1hi} = n_{1h}/N_{1h}$ within each stratum h . Suppose selected PSU i in stratum h is comprised of N_{2hi} secondary sampling units (SSU's). At the second-stage, a SRSWOR of size n_{2hi} SSU's are selected with selection probabilities $\pi_{2hij} = n_{2hi}/N_{2hi}$ within each selected PSU. Suppose selected SSU j in selected PSU i in stratum h is comprised of N_{3hij} ultimate sampling units (USU's). At the third-stage, a SRSWOR of size n_{3hij} USU's are selected with selection probabilities $\pi_{3hijk} = n_{3hij}/N_{3hij}$ within each selected SSU.

The objective is to estimate the population total $Y = \sum_{h=1}^H \sum_{i=1}^{N_{1h}} \sum_{j=1}^{N_{2hi}} \sum_{k=1}^{N_{3hij}} y_{hijk}$, where y_{hijk} is the value for the variable of interest y for USU k in SSU j in PSU i in stratum h . An unbiased estimate of Y is given by:

$$\hat{Y} = \sum_{h=1}^H \hat{Y}_h = \sum_{h=1}^H \frac{N_{1h}}{n_{1h}} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}}{n_{2hi}} \sum_{j=1}^{n_{2hi}} \frac{N_{3hij}}{n_{3hij}} \sum_{k=1}^{n_{3hij}} y_{hijk}$$

where $\hat{Y}_h = (N_{1h}/n_{1h}) \sum_{i=1}^{n_{1h}} \hat{Y}_{hi}$, $\hat{Y}_{hi} = (N_{2hi}/n_{2hi}) \sum_{j=1}^{n_{2hi}} \hat{Y}_{hij}$ and $\hat{Y}_{hij} = (N_{3hij}/n_{3hij}) \sum_{k=1}^{n_{3hij}} y_{hijk}$. This estimator can also be written as $\hat{Y} = \sum_{h=1}^H \sum_{i=1}^{n_{1h}} \sum_{j=1}^{n_{2hi}} \sum_{k=1}^{n_{3hij}} w_{hijk} y_{hijk}$, where $w_{hijk} = w_{1hi} w_{2hij} w_{3hijk} = (N_{1h}/n_{1h})(N_{2hi}/n_{2hi})(N_{3hij}/n_{3hij})$ is the sampling weight for USU k in SSU j in PSU i in stratum h .

An unbiased estimate of $\text{Var}(\hat{Y})$ is given by (Särndal, Swensson and Wretman 1992):

$$\begin{aligned} \hat{\text{Var}}(\hat{Y}) &= \sum_{h=1}^H \frac{N_{1h}^2}{n_{1h}} (1 - f_{1h}) s_{1h}^2 \\ &+ \sum_{h=1}^H \frac{N_{1h}}{n_{1h}} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}^2}{n_{2hi}} (1 - f_{2hi}) s_{2hi}^2 \\ &+ \sum_{h=1}^H \frac{N_{1h}}{n_{1h}} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}}{n_{2hi}} \sum_{j=1}^{n_{2hi}} \frac{N_{3hij}^2}{n_{3hij}} (1 - f_{3hij}) s_{3hij}^2 \quad (2.1) \end{aligned}$$

where $f_{1h} = (n_{1h}/N_{1h})$, $f_{2hi} = (n_{2hi}/N_{2hi})$, $f_{3hij} = (n_{3hij}/N_{3hij})$, $\bar{Y}_h = \sum_{i=1}^{n_{1h}} \hat{Y}_{hi}/n_{1h}$, $\bar{Y}_{hi} = \sum_{j=1}^{n_{2hi}} \hat{Y}_{hij}/n_{2hi}$, $\bar{y}_{hij} = \sum_{k=1}^{n_{3hij}} y_{hijk}/n_{3hij}$, $s_{1h}^2 = \sum_{i=1}^{n_{1h}} (\hat{Y}_{hi} - \bar{Y}_h)^2/(n_{1h} - 1)$, $s_{2hi}^2 = \sum_{j=1}^{n_{2hi}} (\hat{Y}_{hij} - \bar{Y}_{hi})^2/(n_{2hi} - 1)$ and $s_{3hij}^2 = \sum_{k=1}^{n_{3hij}} (y_{hijk} - \bar{y}_{hij})^2/(n_{3hij} - 1)$.

3. Rescaled bootstrap for stratified multistage sampling

Rao and Wu (1988) proposed a rescaling of the standard bootstrap method for various sampling designs including stratified sampling. Since the rescaling factors are applied to the survey data values, this method is only applicable to smooth statistics. Rao, Wu and Yue (1992) presented a modification to this rescaled bootstrap method where the rescaling factors are applied to the survey weights, rather than the survey data values. This modified rescaled bootstrap method is equivalent to the original rescaled bootstrap method, but has the added advantage that it is applicable to non-smooth statistics as well as smooth statistics. Kovar, Rao and Wu (1988) showed that when using a bootstrap sample size of $n_h^* = n_h - 1$ the rescaled bootstrap estimator performed well for smooth statistics.

Although bootstrap samples are usually selected with replacement, Chipperfield and Preston (2007) modified the rescaled bootstrap method to the situation where the bootstrap samples are selected without replacement. Under this without replacement rescaled bootstrap method it can be shown that the choice of either $n_h^* = \lfloor n_h/2 \rfloor$ or $n_h^* = \lceil n_h/2 \rceil$ is optimal, where the operators $\lfloor x \rfloor$ and $\lceil x \rceil$ round the argument x down and up respectively to the nearest integer. The choice of $n_h^* = \lfloor n_h/2 \rfloor$ has the desirable property that the bootstrap weights will never be negative.

For simplicity, the case of stratified three-stage sampling is presented, but the proposed procedure can easily be extended to any number of stages. The without replacement rescaled bootstrap procedure for stratified three-stage sampling is as follows:

(a) Draw a simple random sample of n_{1h}^* PSU's without replacement from the n_{1h} PSU's in the sample. Let δ_{1hi} be equal to 1 if PSU i in stratum h is selected and 0 otherwise. Calculate the PSU bootstrap weights:

$$w_{1hi}^* = w_{1hi} \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}}{n_{1h}^*} \delta_{1hi} \right)$$

where $\lambda_{1h} = \sqrt{n_{1h}^*(1 - f_{1h})/(n_{1h} - n_{1h}^*)}$.

(b) Within each of the PSU's in the sample, draw a simple random sample of n_{2hi}^* SSU's without replacement from the n_{2hi} SSU's in the sample. Let δ_{2hij} be equal to 1 if SSU j in PSU i in stratum h is selected and 0 otherwise. Calculate the conditional SSU bootstrap weights:

$$w_{2hij}^* = \frac{w_{1hi}}{w_{1hi}^*} \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}}{n_{1h}^*} \delta_{1hi} - \lambda_{2hi} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} + \lambda_{2hi} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} \frac{n_{2hi}}{n_{2hi}^*} \delta_{2hij} \right)$$

where $\lambda_{2hi} = \sqrt{n_{2hi}^* f_{1h} (1 - f_{2hi}) / (n_{2hi} - n_{2hi}^*)}$.

(c) Within each of the SSU's in the sample, draw a simple random sample of n_{3hij}^* USU's without replacement from the n_{3hij} USU's in the sample. Let δ_{3hijk} be equal to 1 if USU k in SSU j in PSU i in stratum h is selected and 0 otherwise. Calculate the conditional USU bootstrap weights:

$$w_{3hijk}^* = \frac{w_{1hi}}{w_{1hi}^*} \frac{w_{2hij}}{w_{2hij}^*} \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}}{n_{1h}^*} \delta_{1hi} - \lambda_{2hi} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} + \lambda_{2hi} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} \frac{n_{2hi}}{n_{2hi}^*} \delta_{2hij} - \lambda_{3hij} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} \sqrt{\frac{n_{2hi}}{n_{2hi}^*}} \delta_{2hij} + \lambda_{3hij} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} \sqrt{\frac{n_{2hi}}{n_{2hi}^*}} \delta_{2hij} \frac{n_{3hij}}{n_{3hij}^*} \delta_{3hijk} \right)$$

where $\lambda_{3hij} = \sqrt{n_{3hij}^* f_{1h} f_{2hi} (1 - f_{3hij}) / (n_{3hij} - n_{3hij}^*)}$.

(d) Calculate the bootstrap estimates:

$$\hat{Y}^* = \sum_{h=1}^H \sum_{i=1}^{n_{1h}} \sum_{j=1}^{n_{2hi}} \sum_{k=1}^{n_{3hij}} w_{hijk}^* y_{hijk}, \quad \hat{\theta} = g(\hat{Y}^*)$$

where $w_{hijk}^* = w_{1hi}^* w_{2hij}^* w_{3hijk}^*$.

(e) Independently repeat steps (a) to (d) a large number of times, B , and calculate the bootstrap estimates, $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$, ..., $\hat{\theta}^{(B)}$.

(f) The bootstrap variance estimator of $\hat{\theta}$ is given by:

$$\text{Var}(\hat{\theta}) = E_*(\hat{\theta} - E_*(\hat{\theta}))^2 \quad (3.1)$$

or the Monte Carlo approximation:

$$\text{Var}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2$$

where $\bar{\hat{\theta}} = \sum_{b=1}^B \hat{\theta}^{(b)} / B$.

It is shown in the Appendix that the multistage rescaled bootstrap variance estimator for stratified three-stage sampling as defined by (3.1) reduces to the standard unbiased three-stage variance estimator (2.1) in the case of $\hat{\theta}$ being a linear estimator. The choice of $n_{1h}^* = \lfloor n_{1h} / 2 \rfloor$,

$n_{2hi}^* = \lfloor n_{2hi} / 2 \rfloor$ and $n_{3hij}^* = \lfloor n_{3hij} / 2 \rfloor$ will be optimal and will have the desirable property that the bootstrap weights will never be negative.

The proposed procedure can easily be extended to any number of stages by adding terms of the form $-\lambda_R (\prod_{r=1}^{R-1} \sqrt{(n_r / n_r^*)} \delta_r) + \lambda_R (\prod_{r=1}^{R-1} \sqrt{(n_r / n_r^*)} \delta_r) (n_R / n_R^*) \delta_R$ at each stage, R , to the bootstrap weight adjustments, where $\lambda_R = \sqrt{n_R^* (\prod_{r=1}^{R-1} f_r) (1 - f_R) / (n_R - n_R^*)}$.

Yeo, Mantel and Liu (1999) presented an enhancement to the rescaled bootstrap which accounted for adjustments made to the design weights, such as post-stratification. For example, consider a simple case of non-integrated calibration using auxiliary information for two-stage stratified sampling (Estevao and Särndal 2006), which has the dual objectives of producing estimates for both a first-stage variable of interest $Y_1 = \sum_{(hi) \in U} Y_{1hi}$ as well as a second-stage variable of interest, $Y_2 = \sum_{(hij) \in U} Y_{2hij}$. Assume there exists:

(i) a set of p first-stage auxiliary variables \mathbf{x}_{1hi} for which the population totals $\mathbf{X}_1 = \sum_{(hi) \in U} \mathbf{x}_{1hi}$ are known, and where the population totals are generated from a list frame of PSU's for which the \mathbf{x}_{1hi} are known for every PSU in the population; and

(ii) a set of q second-stage auxiliary variables \mathbf{x}_{2hij} for which the population totals $\mathbf{X}_2 = \sum_{(hij) \in U} \mathbf{x}_{2hij}$ are known, where the population totals are acquired from an external source.

The auxiliary variables can be used to form the first-stage and second-stage calibration estimators:

$$\hat{Y}_{\text{CAL1}} = \sum_{(hi) \in s_1} \tilde{w}_{1hi} y_{1hi}$$

$$\hat{Y}_{\text{CAL2}} = \sum_{(hij) \in s_2} \tilde{w}_{12hij} y_{2hij}$$

where the first-stage calibration weights, \tilde{w}_{1hi} , and the combined first-stage and second-stage calibration weights, \tilde{w}_{12hij} , are given by:

$$\tilde{w}_{1hi} = w_{1hi} \left(1 + \left(\mathbf{X}_1 - \sum_{(hi) \in s_1} w_{1hi} \mathbf{x}_{1hi} \right)^T \left(\sum_{(hi) \in s_1} w_{1hi} \mathbf{x}_{1hi} \mathbf{x}_{1hi}^T \right)^{-1} \mathbf{x}_{1hi} \right)$$

$$\tilde{w}_{12hij} = w_{1hi} w_{2hij} \left(1 + \left(\mathbf{X}_2 - \sum_{(hij) \in s_2} w_{1hi} w_{2hij} \mathbf{x}_{2hij} \right)^T \left(\sum_{(hij) \in s_2} w_{1hi} w_{2hij} \mathbf{x}_{2hij} \mathbf{x}_{2hij}^T \right)^{-1} \mathbf{x}_{2hij} \right)$$

Then the multistage rescaled bootstrap method can easily be modified in a similar manner to handle these calibration estimators by replacing step (d) in the procedure as follows:

(d) Calculate the first-stage and second-stage calibrated bootstrap weights in the same manner as the first-stage and second-stage calibrated weights:

$$\begin{aligned}\tilde{w}_{1hi}^* &= w_{1hi}^* \left(1 + \left(X_1 - \sum_{(hi) \in s_1} w_{1hi}^* x_{1hi} \right)^T \right. \\ &\quad \left. \left(\sum_{(hi) \in s_1} w_{1hi}^* x_{1hi} x_{1hi}^T \right)^{-1} x_{1hi} \right) \\ \tilde{w}_{12hij}^* &= w_{1hi}^* w_{2hij}^* \left(1 + \left(X_2 - \sum_{(hij) \in s_2} w_{1hi}^* w_{2hij}^* x_{2hij} \right)^T \right. \\ &\quad \left. \left(\sum_{(hij) \in s_2} w_{1hi}^* w_{2hij}^* x_{2hij} x_{2hij}^T \right)^{-1} x_{2hij} \right).\end{aligned}$$

The first-stage and second-stage calibrated bootstrap estimates are calculated as:

$$\begin{aligned}\hat{Y}_{CAL1}^* &= \sum_{(hi) \in s_1} \tilde{w}_{1hi}^* y_{1hi} \\ \hat{Y}_{CAL2}^* &= \sum_{(hij) \in s_2} \tilde{w}_{12hij}^* y_{2hij}.\end{aligned}$$

This procedure can easily be modified to any type of calibration and extended to any number of stages. This modification of the rescaled bootstrap takes into account adjustments made to the design weights due to calibration. Ideally all adjustments made to the design weights, including adjustments due to provider non-response and population under-coverage should also be made to the bootstrap weights.

4. Simulation study

A Monte Carlo simulation study was conducted to examine the performance of the multistage rescaled bootstrap variance estimator. The study was restricted to stratified two-stage sampling. The simulation study was based on ten artificial populations, each of which was stratified into $H = 5$ strata, with $N_{1h} = 50$ first-stage units within each stratum, and $N_{2hi} = 40$ second-stage units within each first-stage unit.

Firstly, the first-stage auxiliary variable x_{1hi} for each first-stage unit i in stratum h was generated from the normal distribution $N(\mu_{x1h}, (1 - \rho_{x1b}) \sigma_{x1b}^2 / \rho_{x1b})$. Secondly, the second-stage auxiliary variable, x_{2hij} , and the

second-stage target variables, y_{2hij} and z_{2hij} , for each second-stage unit j within first-stage unit i in stratum h were then generated from the multivariate normal distribution $N_3(\mu_{2hi}, \Sigma_{2hi})$ where μ_{2hi} is the mean vector:

$$\mu_{2hi} = \begin{bmatrix} \mu_{x2hi} \\ \mu_{y2hi} \\ \mu_{z2hi} \end{bmatrix}$$

with $\mu_{x2hi} = \mu_{y2hi} = \mu_{z2hi} = x_{1hi}$, and Σ_{2hi} is the variance-covariance matrix:

$$\Sigma_{2hi} = \begin{bmatrix} \sigma_{x2hi}^2 & \rho_{xy2hi} \sigma_{x2hi} \sigma_{y2hi} & \rho_{xz2hi} \sigma_{x2hi} \sigma_{z2hi} \\ \rho_{xy2hi} \sigma_{x2hi} \sigma_{y2hi} & \sigma_{y2hi}^2 & \rho_{yz2hi} \sigma_{y2hi} \sigma_{z2hi} \\ \rho_{xz2hi} \sigma_{x2hi} \sigma_{z2hi} & \rho_{yz2hi} \sigma_{y2hi} \sigma_{z2hi} & \sigma_{z2hi}^2 \end{bmatrix}$$

with $\sigma_{x2hi}^2 = \sigma_{y2hi}^2 = \sigma_{z2hi}^2 = (1 - \rho_{w2hi}) \sigma_{w2hi}^2 / \rho_{w2hi}$.

The parameter values that were kept stable across all ten populations were $\mu_{x1h} = 25 \times (h + 1)$, $\sigma_{b1h}^2 = 10$, $\sigma_{w2hi}^2 = 100$, $\rho_{xy2hi} = \rho_{xz2hi} = 0.75$ and $\rho_{yz2hi} = 0.50$. The parameter values that were varied across the ten populations were f_{1h} , the first-stage sampling fractions, f_{2hi} , the second-stage sampling fractions, ρ_{b1h} and ρ_{w2hi} . These parameter values are presented in Table 1.

Table 1
Characteristics of simulation populations

	f_{1h}	f_{2hi}	ρ_b	ρ_w
Pop I	0.1	0.1	0.75	0.75
Pop II	0.1	0.1	0.25	0.75
Pop III	0.1	0.5	0.75	0.75
Pop IV	0.1	0.5	0.25	0.75
Pop V	0.1	0.5	0.25	0.25
Pop VI	0.5	0.1	0.75	0.75
Pop VII	0.5	0.1	0.75	0.25
Pop VIII	0.5	0.1	0.25	0.25
Pop IX	0.3	0.3	0.75	0.25
Pop X	0.3	0.3	0.25	0.25

The parameters of interest used in the simulation study were the population mean, μ_y , the population ratio, $R_{yz} = \mu_y / \mu_z$, the population correlation coefficient, $\rho_{yz} = \sigma_{yz} / \sigma_y \sigma_z$, the population regression coefficient, $\beta_{yz} = \sigma_{yz} / \sigma_y^2$, and the population median, M_y .

In order to estimate these parameters of interest using the multistage bootstrap variance estimators, a total of $S = 20,000$ independent two-stage simple random samples were selected without replacement from each of the ten artificial populations. In addition, a grand total of $T = 100,000$ independent two-stage simple random samples were selected without replacement from each of the ten artificial

populations in order to estimate the true population variances for the parameters of interest. The multistage bootstrap variance estimators were calculated using $B = 100$ bootstrap samples.

The accuracy of the multistage bootstrap variance estimators were compared using the relative biases (RB) and the relative root mean square error (RRMSE). These measures were calculated as:

$$RB = \frac{1}{\hat{\text{Var}}(\hat{Y})} \left[\frac{1}{S} \sum_{s=1}^S (\text{Var}_s(\hat{Y}_s) - \hat{\text{Var}}(\hat{Y})) \right]$$
$$RRMSE = \frac{1}{\hat{\text{Var}}(\hat{Y})} \sqrt{\frac{1}{S} \sum_{s=1}^S (\text{Var}_s(\hat{Y}_s) - \hat{\text{Var}}(\hat{Y}))^2}$$

where $\hat{\text{Var}}(\hat{Y}) = T^{-1} \sum_{i=1}^T (\hat{Y}_i - Y)^2$ is the estimated true population variance, and $\text{Var}_s(\hat{Y}_s)$ are the multistage bootstrap variance estimators for the s^{th} simulation sample.

The multistage rescaled bootstrap variance estimator (MRBE) was compared against the single-stage rescaled bootstrap variance estimator (SRBE) and the multistage general Bernoulli bootstrap variance estimator (BBE) proposed by Funaoka *et al.* (2006), with bootstrap samples using the non-calibration estimation weights, $w_{hij} =$

$w_{1hi}w_{2hij}$. The relative biases and relative root mean square errors of MRBE, SRBE and BBE using the non-calibration estimation weights for the ten artificial populations are given in Tables 2 and 3.

In the case of linear functions, such as means, and non-linear functions, such as ratios, correlation coefficients and regression coefficients, the MRBE performed better than the SRBE and BBE with respect to relative bias and relative root mean square error. While the MRBE performed consistently well across all ten artificial populations, the SRBE only performed well for artificial populations III, IV and V, where the first-stage sampling fractions were small ($f_{1h} = 0.1$) and the second-stage sampling fractions were large ($f_{2hi} = 0.5$), and the BBE only performed well for artificial populations VI, VII and VIII, where the first-stage sampling fractions were large ($f_{1h} = 0.5$) and the second-stage sampling fractions were small ($f_{2hi} = 0.1$). These sampling fractions were similar to the first-stage and second-stage sampling fractions used in the simulation study presented in Funaoka *et al.* (2006). The different levels of correlation between the first-stage units, and between the second-stage units within the first-stage units, controlled by varying the parameters ρ_b and ρ_w , had little impact on the performance of the variance estimators.

Table 2
Relative bias (%) of variance estimators

	Mean (μ_y)			Mean (μ_z)			Ratio (R_{yz})		
	MRBE	SRBE	BBE	MRBE	SRBE	BBE	MRBE	SRBE	BBE
Pop I	-0.28	-6.73	27.10	0.42	-6.63	27.32	0.00	-9.07	36.22
Pop II	-0.05	-2.21	11.83	0.59	-1.64	12.54	-0.43	-9.26	36.40
Pop III	-0.79	-2.63	3.62	-0.93	-2.66	3.40	-0.17	-5.30	5.19
Pop IV	-0.23	-0.52	3.60	-0.18	-0.46	3.61	0.53	-4.65	5.98
Pop V	0.15	-1.60	4.55	0.15	-1.64	4.54	0.52	-4.85	6.41
Pop VI	0.70	-39.18	-0.34	0.65	-39.36	-0.28	1.57	-46.40	1.30
Pop VII	0.19	-46.19	-0.26	-0.06	-46.48	-0.57	-0.27	-48.19	-0.73
Pop VIII	0.37	-38.62	-0.41	0.23	-39.36	-0.46	-0.26	-47.93	-0.62
Pop IX	0.42	-20.85	-7.76	-0.51	-20.03	-8.41	0.13	-23.13	-8.87
Pop X	-0.56	-12.35	-6.08	0.70	-10.87	-6.93	-0.72	-23.70	-9.51
	Correlation Coefficient (ρ_{yz})			Regression Coefficient (β_{yz})			Median (M_y)		
	MRBE	SRBE	BBE	MRBE	SRBE	BBE	MRBE	SRBE	BBE
Pop I	-2.31	-10.23	32.17	-0.08	-9.05	36.41	19.04	-19.86	33.21
Pop II	-1.51	-8.41	29.65	0.05	-8.74	36.41	19.29	2.42	40.85
Pop III	0.36	-4.37	5.69	0.05	-5.12	5.42	7.50	4.28	9.72
Pop IV	2.18	-0.60	7.17	0.28	-5.05	5.70	17.40	16.17	34.37
Pop V	0.79	-2.71	5.95	0.26	-5.40	6.34	8.29	4.78	11.49
Pop VI	0.32	-46.67	0.14	0.89	-46.59	0.69	13.57	-33.56	9.15
Pop VII	-0.07	-46.78	-0.39	-0.21	-47.85	-0.60	14.68	-38.16	11.86
Pop VIII	0.31	-44.25	-0.27	-0.09	-47.54	-0.55	2.09	-38.90	-0.64
Pop IX	-0.93	-23.02	-9.30	-0.20	-23.48	-9.20	8.08	-17.23	-1.97
Pop X	-0.82	-19.35	-8.24	-1.02	-23.89	-9.75	2.10	-13.84	-5.46

Note: The largest simulation error on the relative biases was less than 0.7%.

In the case of non-smooth statistics, such as medians, both the MRBE and the BBE tended to overestimate the true population variances, while the SRBE tended to underestimate the true population variances. Furthermore, the relative root mean square errors for medians were up to 3 times larger than the relative root mean square errors for means. The MRBE performed better than the BBE for the artificial populations I to V where the first-stage sampling fractions were smaller ($f_{1h} = 0.1$), while the BBE performed slightly better than the MRBE for the artificial populations VI to X where the first-stage sampling fractions were larger ($f_{1h} = 0.3$ or 0.5).

This overestimation of the multistage rescaled bootstrap for medians was similar to the findings shown in the

studies by Kovar *et al.* (1988) and Rao *et al.* (1992) for the single-stage rescaled bootstrap. It should be noted that the original rescaled bootstrap introduced by Rao and Wu (1988) was developed only for smooth statistics, such as means, ratios, and correlation and regression coefficients.

The MRBE was examined using the calibration estimation weights, $\bar{w}_{hij} = w_{1hi}\bar{w}_{2hij}$, which satisfy the calibration constraint $\sum_{(hij) \in s_2} w_{1hi}\bar{w}_{2hij} = X_2$, where $X_2 = \sum_{(hij) \in U} x_{2hij}$ is the population total for the second-stage auxiliary variable. The relative biases and relative root mean square errors of the MRBE using the calibration estimation weights for the four artificial populations II, IV, VII and IX are given in Table 4.

Table 3
Relative root mean square error (%) of variance estimators

	Mean (μ_y)			Mean (μ_z)			Ratio (R_{yz})		
	MRBE	SRBE	BBE	MRBE	SRBE	BBE	MRBE	SRBE	BBE
Pop I	31.9	32.1	44.6	31.7	31.8	44.4	31.8	32.3	51.4
Pop II	33.9	33.8	38.2	33.4	33.3	38.1	32.2	32.9	51.7
Pop III	33.8	33.8	35.9	33.0	33.0	35.0	33.0	33.1	35.1
Pop IV	35.3	35.3	37.4	35.2	35.2	37.3	32.8	32.8	35.0
Pop V	32.0	31.9	34.2	34.3	34.2	36.5	33.0	33.1	35.6
Pop VI	16.4	40.7	16.5	16.4	40.9	16.5	16.5	47.5	16.5
Pop VII	16.1	47.4	16.4	16.1	47.8	16.4	16.1	49.0	16.1
Pop VIII	16.3	40.3	16.5	16.7	40.9	16.3	16.2	48.8	16.1
Pop IX	19.2	26.7	20.0	19.3	26.3	20.0	19.2	28.6	20.2
Pop X	19.8	22.4	20.2	19.9	21.6	20.3	19.1	29.0	20.6

	Correlation Coefficient (ρ_{yz})			Regression Coefficient (β_{yz})			Median (M_y)		
	MRBE	SRBE	BBE	MRBE	SRBE	BBE	MRBE	SRBE	BBE
Pop I	47.8	46.3	68.7	36.6	37.2	55.3	88.7	80.1	89.8
Pop II	48.4	47.1	66.6	37.4	37.9	55.6	93.4	91.0	115.9
Pop III	35.9	35.6	38.4	37.5	37.6	39.9	80.4	80.3	81.1
Pop IV	42.6	42.2	45.4	38.0	38.0	40.3	97.5	96.6	127.3
Pop V	40.3	40.0	43.3	37.3	37.5	40.1	31.5	30.7	63.3
Pop VI	21.6	48.4	21.7	16.9	47.8	17.0	55.3	51.4	52.0
Pop VII	21.4	48.4	21.3	16.9	49.0	16.8	53.5	51.4	51.4
Pop VIII	21.6	46.3	21.5	17.0	48.6	16.9	41.8	49.7	40.3
Pop IX	21.5	29.4	22.5	20.5	29.9	21.6	46.1	42.7	42.7
Pop X	22.7	27.8	23.4	20.6	30.2	21.9	39.7	38.9	37.9

Table 4
Relative bias (%) and relative root mean square error (%) of rescaled bootstrap variance estimator

	μ_y	R_{yz}	ρ_{yz}	β_{yz}	M_y
	Relative Bias (%)				
Pop II	-0.42	-0.29	-1.51	-0.08	20.98
Pop IV	0.40	0.49	1.83	0.08	18.28
Pop VII	-0.22	-0.24	-0.03	-0.28	12.24
Pop IX	0.62	0.19	-1.00	-0.20	7.24
	Relative Root Mean Square Error (%)				
Pop II	32.6	32.4	48.4	37.3	97.8
Pop IV	32.8	32.8	44.6	37.9	99.4
Pop VII	16.2	16.1	21.5	16.9	50.0
Pop IX	19.1	19.2	21.6	20.5	43.8

Note: The largest simulation error on the relative biases was less than 0.6%.

The relative biases and relative root mean square errors of the MRBE using the calibration estimation weights were similar to those using the non-calibration estimation weights.

5. Conclusion

This paper extends the rescaled bootstrap procedure to multistage sampling where units are selected using simple random sampling without replacement at each stage. Under the proposed multistage rescaled bootstrap method, the bootstrap samples are selected without replacement and rescaling factors are applied to the survey weights. This proposed method is relatively simple to implement and requires considerably less random number generations than the multistage general Bernoulli bootstrap method. The proposed method is also suitable for a wide range of reweighting techniques, including calibration, and adjustments due to provider non-response and population under-coverage. Furthermore, the results of the Monte Carlo simulation study indicate that the multistage rescaled bootstrap performs much better than the single-stage rescaled bootstrap and the multistage Bernoulli bootstrap for smooth statistics, such as means, ratios, and correlation and regression coefficients.

Appendix

In this Appendix it is shown that the multistage rescaled bootstrap variance estimator for stratified three-stage sampling reduces to the standard unbiased three-stage variance estimator (2.1) in the case of $\hat{\theta}$ being the linear estimator, $\hat{Y}^* = \sum_{h=1}^H \sum_{i=1}^{n_{1h}} \sum_{j=1}^{n_{2hi}} \sum_{k=1}^{n_{3hij}} w_{hijk}^* y_{hijk}$. The bootstrap variance estimator of \hat{Y}^* is given by:

$$\begin{aligned} \text{Var}(\hat{Y}^*) &= \text{Var}_*(E_{2*}(E_{3*}(\hat{Y}^*))) \\ &+ E_{1*}(\text{Var}_{2*}(E_{3*}(\hat{Y}^*))) + E_{1*}(E_{2*}(\text{Var}_{3*}(\hat{Y}^*))). \end{aligned}$$

Using standard results on the expectation and variance with respect to the SRSWOR bootstrap sampling and some tedious but straightforward algebra, the components of bootstrap variance estimator are given below. The conditional expectation of \hat{Y}^* given s_3 is

$$\begin{aligned} E_{3*}(\hat{Y}^*) &= \sum_{h=1}^H \sum_{i=1}^{n_{1h}} \sum_{j=1}^{n_{2hi}} w_{1i} w_{2ij} \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}}{n_{1h}^*} \delta_{1hi} \right. \\ &\quad \left. - \lambda_{2hi} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \delta_{1hi} + \lambda_{2hi} \sqrt{\frac{n_{1h}}{n_{1h}^*}} \frac{n_{2hi}}{n_{2hi}^*} \delta_{2hij} \right) \hat{Y}_{hi} \end{aligned}$$

and the conditional variance of \hat{Y}^* given s_3 is

$$\begin{aligned} \text{Var}_{3*}(\hat{Y}^*) &= \sum_{h=1}^H \frac{N_{1h}}{n_{1h}^*} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}}{n_{2hi}^*} \sum_{j=1}^{n_{2hi}} \delta_{1hi} \delta_{2hij} \frac{N_{3hij}^2}{n_{3hij}} (1 - f_{3hij}) s_{3hij}^2. \end{aligned}$$

The conditional expectation of $E_{3*}(\hat{Y}^*)$ and $\text{Var}_{3*}(\hat{Y}^*)$ given s_2 are

$$E_{2*}(E_{3*}(\hat{Y}^*)) = \sum_{h=1}^H \sum_{i=1}^{n_{1h}} w_{1i} (1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}}{n_{1h}^*} \delta_{1hi}) \hat{Y}_{hi}$$

$$\begin{aligned} E_{2*}(\text{Var}_{3*}(\hat{Y}^*)) &= \sum_{h=1}^H \frac{N_{1h}}{n_{1h}^*} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}}{n_{2hi}^*} \sum_{j=1}^{n_{2hi}} \delta_{1hi} \frac{N_{3hij}^2}{n_{3hij}} (1 - f_{3hij}) s_{3hij}^2 \end{aligned}$$

and the conditional variance of $E_{3*}(\hat{Y}^*)$ given s_2 is

$$\text{Var}_{2*}(E_{3*}(\hat{Y}^*)) = \sum_{h=1}^H \frac{N_{1h}}{n_{1h}^*} \sum_{i=1}^{n_{1h}} \delta_{1hi} \frac{N_{2hi}^2}{n_{2hi}^*} (1 - f_{2hi}) s_{2hi}^2.$$

Finally, the conditional expectation of $E_{2*}(\text{Var}_{3*}(\hat{Y}^*))$ and $\text{Var}_{2*}(E_{3*}(\hat{Y}^*))$ given s_1 are

$$\begin{aligned} E_{1*}(E_{2*}(\text{Var}_{3*}(\hat{Y}^*))) &= \sum_{h=1}^H \frac{N_{1h}}{n_{1h}^*} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}}{n_{2hi}^*} \sum_{j=1}^{n_{2hi}} \frac{N_{3hij}^2}{n_{3hij}} (1 - f_{3hij}) s_{3hij}^2 \end{aligned}$$

$$E_{1*}(\text{Var}_{2*}(E_{3*}(\hat{Y}^*))) = \sum_{h=1}^H \frac{N_{1h}}{n_{1h}^*} \sum_{i=1}^{n_{1h}} \frac{N_{2hi}^2}{n_{2hi}^*} (1 - f_{2hi}) s_{2hi}^2$$

which are equal to the third and second terms of (2.1) respectively, and the conditional variance of $E_{2*}(E_{3*}(\hat{Y}^*))$ given s_1 is

$$\text{Var}_{1*}(E_{2*}(E_{3*}(\hat{Y}^*))) = \sum_{h=1}^H \frac{N_{1h}^2}{n_{1h}^*} (1 - f_{1h}) s_{1h}^2$$

which is equal to the first term of (2.1).

References

- Australian Bureau of Statistics (ABS) (2008). Employee Earnings and Hours, Catalogue Number 6306.0.
- Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, 12, 470-482.
- Chipperfield, J., and Preston, J. (2007). Efficient bootstrap for business surveys. *Survey Methodology*, 33, 167-172.

- Estevao, V., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.
- Funaoka, F., Saigo, H., Sitter, R.R. and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32, 151-156.
- Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, 181-184.
- Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25-45.
- McCarthy, P.J., and Snowden, C.B. (1985). The bootstrap and finite population sampling. Vital and Health Statistics (Series 2 No 95), Public Health Service Publication 85-1369, Washington, DC: U.S. Government Printing Office.
- Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- Särndal, C.-E., Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.
- Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.
- White, V., and Hayman J. (2006). Smoking behaviours of Australian secondary students in 2005. National Drug Strategy Monograph Series No. 59. Canberra: Australian Government Department of Health and Ageing.
- Yeo, D., Mantel, H. and Liu T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

Use of within-primary-sample-unit variances to assess the stability of a standard design-based variance estimator

Donsig Jang and John L. Eltinge¹

Abstract

In analysis of sample survey data, degrees-of-freedom quantities are often used to assess the stability of design-based variance estimators. For example, these degrees-of-freedom values are used in construction of confidence intervals based on t distribution approximations; and of related t tests. In addition, a small degrees-of-freedom term provides a qualitative indication of the possible limitations of a given variance estimator in a specific application. Degrees-of-freedom calculations sometimes are based on forms of the Satterthwaite approximation. These Satterthwaite-based calculations depend primarily on the relative magnitudes of stratum-level variances. However, for designs involving a small number of primary units selected per stratum, standard stratum-level variance estimators provide limited information on the true stratum variances. For such cases, customary Satterthwaite-based calculations can be problematic, especially in analyses for subpopulations that are concentrated in a relatively small number of strata. To address this problem, this paper uses estimated within-primary-sample-unit (within PSU) variances to provide auxiliary information regarding the relative magnitudes of the overall stratum-level variances. Analytic results indicate that the resulting degrees-of-freedom estimator will be better than modified Satterthwaite-type estimators provided: (a) the overall stratum-level variances are approximately proportional to the corresponding within-stratum variances; and (b) the variances of the within-PSU variance estimators are relatively small. In addition, this paper develops errors-in-variables methods that can be used to check conditions (a) and (b) empirically. For these model checks, we develop simulation-based reference distributions, which differ substantially from reference distributions based on customary large-sample normal approximations. The proposed methods are applied to four variables from the U.S. Third National Health and Nutrition Examination Survey (NHANES III).

Key Words: Complex sample design; Degrees of freedom; Errors-in-variables regression; Satterthwaite approximation; Stratified multistage sample survey; Two-PSU-per-stratum design; U.S. Third National Health and Nutritional Examination Survey (NHANES III).

1. Introduction

1.1 Motivating example: Inference for special subpopulations in NHANES III

This work arose from a study of inference for geographically concentrated subpopulations in the U.S. Third National Health and Nutrition Examination Survey (NHANES III). For some general background on NHANES III, see National Center for Health Statistics (1996). In many analyses, NHANES III data are treated as arising from a stratified multistage sample design that uses 49 strata and two primary sample units (PSUs) per stratum. Consequently, formal inferences from NHANES III data (e.g., construction of confidence intervals) often use the assumption that the associated variance estimators are based on approximately 49 degrees of freedom and are thus relatively stable.

However, the Mexican-American subpopulation is concentrated in a relatively small number of strata, so associated variance estimators may be less stable (i.e., have greater sampling variability) than would be indicated by the nominal 49 degrees of freedom term. Consequently, it is important to use an appropriate estimator of the true degrees of freedom associated with variance estimators for such

subpopulations, and to modify confidence interval calculations accordingly. Development of an appropriate degrees-of-freedom estimator can be complicated by moderate or severe heterogeneity in the underlying stratum-level variances. Such complications arose in the analysis of the four NHANES III variables listed in Table 1.1. Section 5 will consider inference for the means of these four variables for the subpopulation of Mexican-Americans aged 20-29.

Table 1.1
Four NHANES III variables

Variable Name	Description
BMPWT	Weight (kg)
HAR3	Do you smoke cigarettes now? (0/1)
TCRESULT	Serum total cholesterol (mg/dL)
HDRESULT	HDL cholesterol (mg/dL)

1.2 Stability of design-based variance estimators

Suppose we have a population partitioned into L strata, with N_h PSUs in stratum h for $h = 1, 2, \dots, L$. Under a

1. Donsig Jang, Mathematica Policy Research, 600 Maryland Avenue SW, Suite 550, Washington, DC 20024-2512, U.S.A. E-mail: DJang@Mathematica-mpr.com; John L Eltinge, U.S. Bureau of Labor Statistics, PSB 1950, 2 Massachusetts Avenue NE, Washington, DC 20212-0001, U.S.A. E-mail: Eltinge_J@bls.gov.

stratified multistage sampling design, we select n_h PSUs, with replacement, and with per-draw selection probability p_{hi} for PSU i within stratum h where $\sum_{i=1}^{N_{hi}} p_{hi} = 1$. Thus, a total of $n = \sum_{h=1}^L n_h$ PSUs are selected. Within selected PSU (h, i) , n_{hi} secondary sample units (SSUs) are selected with replacement and with per-draw selection probabilities p_{hij} , where $\sum_{j=1}^{N_{hij}} p_{hij} = 1$ and N_{hi} is the number of SSUs in PSU (h, i) . For a given survey item, let Y_h be the population total for stratum h , and define the overall population total $Y = \sum_{h=1}^L Y_h$. The total Y may correspond to a total either for the full population or for a specified subpopulation.

Our goal is to construct a confidence interval for the total Y . Let \hat{Y}_{hij} be an unbiased estimator of Y_{hij} , the population total for secondary unit j in primary unit i in stratum h . Then a customary design-based estimator of Y is $\hat{Y} = \sum_{h=1}^L \hat{Y}_h$, where $\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \hat{Y}_{hi}$; $p_{hi}^{-1} \hat{Y}_{hi}$ is a design unbiased estimator of Y_{hi} based on data obtained from PSU i in stratum h ; and $\hat{Y}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} p_{hij}^{-1} \hat{Y}_{hij}$ is an unbiased estimator of Y_{hi} , the population total for PSU i in stratum h .

Under the standard condition that sampling is independent across strata, the variance of \hat{Y} can be written, $V(\hat{Y}) = \sum_{h=1}^L V_h$ where $V_h = \text{Var}(\hat{Y}_h)$. Throughout the remainder of this paper, we will call the V_h terms the stratum-level variances, and we will assume that $n_h \geq 2$ for all $h = 1, 2, \dots, L$. Note that V_h depends on the sample design used within stratum h , and is distinct from the within-stratum variance of element-level Y values. A simple unbiased estimator for $V(\hat{Y})$ is $\hat{V}(\hat{Y}) = \sum_{h=1}^L \hat{V}_h$ where $\hat{V}_h = n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (p_{hi}^{-1} \hat{Y}_{hi} - \hat{Y}_h)^2$; see, e.g., Wolter (1985, page 44). Note that the estimator \hat{V}_h is a multiple of a sum of squared differences among the terms $p_{hi}^{-1} \hat{Y}_{hi}$. In addition, under regularity conditions the random variables $p_{hi}^{-1} \hat{Y}_{hi}$ will be approximately normally distributed for a given stratum h . Consequently, the overall stratum-level variance estimators \hat{V}_h generally will approximately satisfy the following condition.

(C.1) For $h = 1, 2, \dots, L$, the terms $V_h^{-1} (n_h - 1) \hat{V}_h$ are distributed as independent chi-square random variables with $n_h - 1$ degrees of freedom, respectively, where $n_h \geq 2$.

Under condition (C.1), $\{V(\hat{Y})\}^{-1} d \hat{V}(\hat{Y})$ has the same first and second moments as a chi-square random variable with d degrees of freedom, where d is the solution to the equation,

$$2\{V(\hat{Y})\}^2 - V\{\hat{V}(\hat{Y})\} d = 0 \quad (1.1)$$

or equivalently

$$d \stackrel{\text{def}}{=} \left\{ \sum_{h=1}^L (n_h - 1)^{-1} V_h^2 \right\}^{-1} \{V(\hat{Y})\}^2 \quad (1.2)$$

where $V\{\hat{V}(\hat{Y})\} = \sum_{h=1}^L 2(n_h - 1)^{-1} V_h^2$. Direct substitution of \hat{V}_h for V_h and $\hat{V}(\hat{Y})$ for $V(\hat{Y})$ in expression (1.2) leads to the Satterthwaite (1946)-type degrees-of-freedom estimator,

$$\hat{d}_S = \left\{ \sum_{h=1}^L (n_h - 1)^{-1} \hat{V}_h^2 \right\}^{-1} \{\hat{V}(\hat{Y})\}^2. \quad (1.3)$$

For some general background on \hat{d}_S and related estimators, see, e.g., Smith (1936), Satterthwaite (1941, 1946), Cochran (1977, page 96) and Kendall, Stuart and Ord (1983, pages 91-92). In constructing confidence intervals for a subpopulation parameter, Casady, Dorfman and Wang (1998) use Bayesian ideas to develop related degrees-of-freedom measures for a Student's t -statistic.

For designs in which n_h is large for all h , the error in estimation of V_h is relatively small, and \hat{d}_S can provide a satisfactory estimator of expression (1.2). However, many large-scale surveys use small n_h , e.g., $n_h = 2$. For small- n_h cases, condition (C.1) and routine algebra lead to the expectation result $E(\hat{V}_h^2) = (n_h - 1)^{-1} (n_h + 1) V_h^2$. This implies that the standard Satterthwaite degrees-of-freedom estimator \hat{d}_S can severely underestimate d , and that the corresponding confidence interval $\hat{Y} \pm t_{\hat{d}_S, 1-\alpha/2} \{\hat{V}(\hat{Y})\}^{1/2}$ may have a true coverage rate substantially below the nominal rate $1 - \alpha$. Consequently, Jang (1996) considered an alternative degrees-of-freedom estimator,

$$\hat{d}_{ms} = (3L + 14)^{-1} (9L) \hat{d}_S. \quad (1.4)$$

for the two-PSUs-per-stratum design.

1.3 Use of auxiliary stratum-level data

For cases in which there is moderate heterogeneity among the V_h terms, simulation work by Jang (1996) indicated that \hat{d}_{ms} performs relatively well. However, if there is substantial heterogeneity among the stratum variances (i.e., if $L^{-1}d$ is relatively small), then \hat{d}_{ms} may be unsatisfactory. The fundamental problem is that when the n_h values are relatively small, the estimators \hat{V}_h , by themselves, do not provide sufficient information regarding the relative magnitudes of the true stratum-level variances V_h . In some cases, a variance estimator based on auxiliary data is expected to be more stable than the customary design-based estimator; see e.g., Isaki (1983). Similarly, auxiliary sources of information can be used to evaluate the relative magnitudes of the variances V_h .

The remainder of this paper will focus on auxiliary information provided by relationships between the overall stratum-level variances V_h and associated within-PSU variances. Recall from Wolter (1985, page 41) the decomposition,

$$\text{Var}(\hat{Y}_h) = V_{Bh} + V_{Wh}, \quad (1.5)$$

where $V_{Bh} = \text{Var}\{\sum_{i=1}^{n_h} (n_h p_{hi})^{-1} Y_{hi}\}$ is the between-PSU variance, $V_{wh} = \sum_{i=1}^{n_h} (n_h p_{hi})^{-1} \sigma_{2hi}^2$ is the within-PSU variance, $Y_{hi} = E(\hat{Y}_{hi} | \text{PSU } i, \text{ stratum } h)$ and $\sigma_{2hi}^2 = \text{Var}(\hat{Y}_{hi} | \text{PSU } i, \text{ stratum } h)$. In addition, define $\bar{V}_w = L^{-1} \sum_{h=1}^L V_{wh}$.

Estimators of V_{wh} can provide useful auxiliary information on the relative magnitudes of V_h for two reasons. First, for designs with a small n_h and relatively large n_{hi} , the within-PSU variance estimators \hat{V}_{wh} may be considerably more stable than \hat{V}_h . Second, in some applications (e.g., some of the examples presented in Section 5 below), observed variance estimates are consistent with a model under which V_h is proportional to V_{wh} , i.e.,

$$V_h = \beta_1 V_{wh} \text{ for all } h = 1, \dots, L, \quad (1.6)$$

where β_1 is a fixed constant. The proportionality relationship (1.6) would arise if both V_{Bh} and V_{wh} are proportional to a common scale factor, e.g., $(\bar{Y}_h)^\alpha$ for some power α . Under relationship (1.6), expression (1.2) may be rewritten,

$$d = \left\{ \sum_{h=1}^L (n_h - 1)^{-1} V_{wh}^2 \right\}^{-1} \left\{ \sum_{h=1}^L V_{wh} \right\}^2. \quad (1.7)$$

Consequently, given a set of stable within-PSU variance estimators \hat{V}_{wh} and associated variance-of-variance-estimators $\widehat{\text{Var}}(\hat{V}_{wh})$,

$$\hat{d}_{wS} = \left\{ \sum_{h=1}^L (n_h - 1)^{-1} [\hat{V}_{wh}^2 - \widehat{\text{Var}}(\hat{V}_{wh})] \right\}^{-1} \left(\sum_{h=1}^L \hat{V}_{wh} \right)^2 \quad (1.8)$$

is an alternative estimator of d .

Section 2 considers some of the properties of \hat{d}_{wS} . Section 3.1 uses errors-in-variables tests to check the adequacy of the proportionality condition (1.6). Section 3.2 presents two related diagnostics for the relationship between V_h and auxiliary variables, and for the magnitude of the error in the observed auxiliary variables \hat{V}_{wh} .

A simulation study in Section 4 explores conditions under which the proposed new estimator \hat{d}_{wS} may perform better than \hat{d}_{mS} . This assessment considers both the estimation of d as such, and the performance of confidence intervals for Y . Section 5 applies the proposed estimator to four variables from NHANES III, with emphasis on cases for which differences between the proposed estimators \hat{d}_{wS} and \hat{d}_{mS} have a substantial practical effect on assessment of the stability of the variance estimator $\hat{V}(\hat{Y})$. Section 6 reviews the methods developed in this paper and considers some possible extensions.

2. An estimator based on auxiliary information

2.1 A within-PSU variance estimator

A simple estimator of V_{wh} is

$$\hat{V}_{wh} = n_h^{-2} \sum_{i=1}^{n_h} p_{hi}^{-2} \hat{\sigma}_{2hi}^2, \quad (2.1)$$

where $\hat{\sigma}_{2hi}^2 = n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (p_{hij}^{-1} \hat{Y}_{hij} - \hat{Y}_{hi})^2$. Note that $\hat{\sigma}_{2hi}^2$ is approximately unbiased for σ_{2hi}^2 under a with-replacement sampling design within PSU i in stratum h ; or under simple random sampling without replacement and with a small sampling fraction, $f_{hi} = N_{hi}^{-1} n_{hi}$. Standard sampling theory shows that \hat{V}_{wh} is approximately unbiased for V_{wh} . Then an approximately unbiased estimator of $\text{Var}(\hat{V}_{wh})$ is

$$\widehat{\text{Var}}(\hat{V}_{wh}) = n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (\hat{V}_{whi} - \hat{V}_{wh})^2, \quad (2.2)$$

where $\hat{V}_{whi} = n_{hi}^{-1} p_{hi}^{-2} \hat{\sigma}_{2hi}^2$; see, e.g., Eltinge and Jang (1996) and references cited therein. Note that the overall stratum-level variance estimators \hat{V}_h are functions of the sample means of $p_{hij}^{-1} \hat{Y}_{hij}$ over PSUs in stratum h . In addition, the estimators \hat{V}_{wh} are functions of sample variances of the $p_{hij}^{-1} \hat{Y}_{hij}$ within the PSU (h, i) . Thus, for variables Y for which $p_{hij}^{-1} \hat{Y}_{hij}$ are approximately normally distributed within stratum h , the estimators \hat{V}_h and \hat{V}_{wh} are approximately independent.

2.2 Properties of \hat{d}_{wS}

In the remainder of this paper, the estimator \hat{d}_{wS} defined in expression (1.8) will use $\widehat{\text{Var}}(\hat{V}_{wh})$ as defined in expression (2.2). Also, the remainder of this paper will use several asymptotic results. These results will use the condition that the number of strata, L , is increasing, while stratum-level PSU and SSU sample sizes n_h and m_h are allowed to remain small. This is in keeping with many practical multi-stage designs that use $n_h = 2$ and moderate values of m_h . See, e.g., Krewski and Rao (1981) for a detailed development of large- L asymptotic results. The proof of Result 2.1 is routine and is thus omitted.

Result 2.1. Assume that $E(\hat{V}_{wh}^r) = O(1)$ for $r = 1, 2, 3, 4$ and define

$$\hat{\bar{V}}_w = L^{-1} \sum_{h=1}^L \hat{V}_{wh} \quad (2.3)$$

and

$$\hat{V}_{w(2)} = L^{-1} \sum_{h=1}^L (n_h - 1)^{-1} \{ \hat{V}_{wh}^2 - \widehat{\text{Var}}(\hat{V}_{wh}) \}.$$

Then \hat{V}_w and $\hat{V}_{w(2)}$ are consistent estimators of \bar{V}_w and $L^{-1} \sum_{h=1}^L (n_h - 1)^{-1} V_{wh}^2$, respectively. In addition, $L^{-1} \hat{d}_{ws}$ is a consistent estimator of $L^{-1} d_{ws}$.

Section 1 suggested that in some cases, the auxiliary-data based estimator \hat{d}_{ws} might be more stable than the modified Satterthwaite estimator \hat{d}_{ms} . To examine this idea, we will compare the variances of \hat{d}_{ws} and \hat{d}_{ms} under condition (C.1) and the following additional assumptions.

(C.2) For $h = 1, 2, \dots, L$, $V_{wh}^{-1}(m_h - 1)\hat{V}_{wh}$ are distributed as independent chi-square random variables with $m_h - 1$ degrees of freedom, respectively, where m_h is the number of SSUs in stratum h ; and are mutually independent of \hat{V}_h .

(C.3) For all $h = 1, 2, \dots, L$, $n_h = 2$; and $m_h = m_0$ for some fixed positive integer $m_0 \geq 2$.

Arguments similar to those for condition (C.1) indicate that condition (C.2) may be satisfied approximately if within a given PSU (h, i), the m_h random variables $p_{hij}^{-1}\hat{Y}_{hij}$ are approximately independent and identically distributed normal random variables. Condition (C.3) restricts attention to the common case $n_h = 2$. In addition, condition (C.3) requires that an equal number, m_0 , of secondary units be selected within each selected PSU. This allows simplification of the resulting approximations for the variances of \hat{d}_{ws} , as presented in Result 2.2.

Result 2.2. Assume conditions (C.1), (C.2), (C.3), and (1.6), and define $a = 4\mu_{A_1}^2 \mu_{B_1}^{-2} \text{Var}(A_2)$, $b = 4\mu_{A_1}^3 \mu_{B_1}^{-3} \text{Cov}(A_2, B_2)$, and $c = \mu_{A_1}^4 \mu_{B_1}^{-4} \text{Var}(B_2)$, where $A_2 = L^{-1} \sum_{h=1}^L \hat{V}_{wh}$, $B_2 = L^{-1} \sum_{h=1}^L \{\hat{V}_{wh}^2 - \text{Var}(\hat{V}_{wh})\}$, $\mu_{A_2} = \bar{V}_w$ and $\mu_{B_2} = L^{-1} \sum_{h=1}^L \hat{V}_{wh}^2$. Then

- (i) the variances of the leading terms in Taylor expansions of $L^{-1}(\hat{d}_{ws} - d)$ and $L^{-1}(\hat{d}_{ms} - d)$ are, respectively,

$$V_{LW} = a - b + c$$

and

$$V_{Lm} = \frac{1}{9} \left(\frac{9L}{3L+14} \right)^2 (m_0 - 1) \left\{ a - b + \frac{4(m_0 - 1)}{3(m_0 + 2)} c \right\}.$$

- (ii) for all $m_0 \geq \lim_{L \rightarrow \infty} g(a, b, c)$, $\lim_{L \rightarrow \infty} V_{Lm} \geq \lim_{L \rightarrow \infty} V_{LW}$ where

$$g(a, b, c) = \{2(3a - 3b + 4c)\}^{-1} \{11c + \sqrt{144a^2 + 144b^2 + 153c^2 - 288ab + 216ac - 216bc}\}.$$

- (iii) for $m_0 \geq 10$, $\lim_{L \rightarrow \infty} V_{Lm} \geq \lim_{L \rightarrow \infty} V_{LW}$ regardless of the values of the limiting moments $\lim_{L \rightarrow \infty} (\mu_{A_2}, \mu_{B_2}, L^{-1} \sum_{h=1}^L V_{wh}^3, L^{-1} \sum_{h=1}^L V_{wh}^4)$.

Result 2.2 indicates that for large L , \hat{d}_{ws} may be preferable to \hat{d}_{ms} , provided: (1) the proportionality condition (1.6) is satisfied; and (2) the secondary unit sample size m_0 exceeds the lower bound given by $g(a, b, c)$ (thus ensuring relatively small variances of the \hat{V}_{wh}). This motivates the use of within-PSU variances to assess the stability of survey variance estimators, especially under sample designs with small numbers of PSUs per stratum. For some additional discussion of this point, and some specific diagnostics to check the stability of \hat{V}_{wh} , see Eltinge and Jang (1996) and references cited therein. For the four cases considered in Table 1.1 and studied further in Section 4 below, $g(a, b, c)$ is equal to 4.7, 4.3, 4.6, and 4.8 respectively, while the NHANES III application had the mean of the m_h values approximately equal to 22. In addition, we are treating V_{wh} values as fixed, and Result 2.2 depends on the limiting moments of these V_{wh} terms. Suppose that V_{wh}/\bar{V}_w had the same moments as F/f , where F follows a chi-square distribution on f degrees of freedom. Then $f = \infty$ corresponds to the case in which $V_{wh} = \bar{V}_w$ for all h , which corresponds to the case in which the true d in (1.1) equals the customary value of $n - L$.

3. Testing the proportionality condition

3.1 An errors-in-variables model for V_h and V_{wh}

Development of the alternative estimator \hat{d}_{ws} in Section 1, and evaluation of its properties in Section 2, depended heavily on the proportionality condition (1.6). One may test the adequacy of this condition through the following steps. First, note that condition (1.6) is a special case of the following model,

(C.4) For all $h = 1, 2, \dots, L$,

$$V_h = \beta_0 + \beta_1 V_{wh} + q_h \quad (3.1)$$

where β_0 and β_1 are constants, and q_h is an equation error with mean zero and variance σ_{qqh} .

Second, recall that V_h and V_{wh} are unknown quantities, for which we have the unbiased estimators \hat{V}_h and \hat{V}_{wh} , respectively. Using the errors-in-variables model notation in Fuller (1987), define the estimation errors

$$e_h = \hat{V}_h - V_h \quad \text{and} \quad u_h = \hat{V}_{wh} - V_{wh}. \quad (3.2)$$

Under conditions (C.1) and (C.2), the vector $(e_h, u_h)'$ is distributed with a mean vector equal to $(0, 0)'$ and a variance-covariance matrix equal to $\text{diag}(\sigma_{eeh}, \sigma_{uuh})$ where $\sigma_{eeh} = (n_h - 1)^{-1} 2V_h^2$ and $\sigma_{uuh} = (m_h - 1)^{-1} 2V_{wh}^2$. Under the additional condition (C.3), these variance terms simplify to $\sigma_{eeh} = 2V_h^2$ and $\sigma_{uuh} = (m_0 - 1)^{-1} 2V_{wh}^2$.

Expressions (3.1) and (3.2) define an errors-in-variables regression model with heterogeneous measurement error variances and non-normal errors. In addition $\widehat{\text{Var}}(\hat{V}_{wh})$ defined in expression (2.2) is an unbiased estimator of σ_{uuh} , and thus provides identifying information for the parameters β_0, β_1 and σ_{qqh} in model (3.1)–(3.2). A direct application of Fuller (1987, pages 187–189) with equal weights then gives the consistent estimators (for increasing L),

$$\begin{aligned}\hat{\beta}_0 &= L^{-1} \sum_{h=1}^L \hat{V}_h - \hat{\beta}_1 \hat{V}_w, \\ \hat{\beta}_1 &= \left[\sum_{h=1}^L (\hat{V}_{wh} - \hat{V}_w)^2 - \hat{\sigma}_{uu} \right]^{-1} \sum_{h=1}^L (\hat{V}_{wh} - \hat{V}_w) \hat{V}_h, \quad (3.3)\end{aligned}$$

and

$$\begin{aligned}\hat{\sigma}_{qq} &= \max \left[0, L^{-1} \sum_{h=1}^L (n_h - 1)^{-1} \right. \\ &\quad \left. \{ (L - 2)^{-1} L (\hat{V}_h - \hat{\beta}_0 - \hat{\beta}_1 \hat{V}_{wh})^2 \right. \\ &\quad \left. - (\hat{\sigma}_{eeh} + \hat{\beta}_1^2 \hat{\sigma}_{uuh}) \right], \quad (3.4)\end{aligned}$$

where

$$\hat{\sigma}_{uu} = \sum_{h=1}^L \widehat{\text{Var}}(\hat{V}_{wh}), \quad \hat{V}_w = L^{-1} \sum_{h=1}^L \hat{V}_{wh}, \quad (3.5)$$

and

$$\hat{\sigma}_{eeh} = 2(n_h + 1)^{-1} \hat{V}_h^2$$

from condition (C.1). In addition, direct application of Fuller (1987, page 188) leads to variance estimators $\hat{V}(\hat{\beta}_0)$ and $\hat{V}(\hat{\beta}_1)$, say; details are available from the authors.

3.2 Two related diagnostics

In keeping with condition (C.4), the proposed estimator \hat{d}_{ws} is intended for cases in which the \hat{V}_{wh} provide useful auxiliary information on the relative magnitudes of the overall stratum-level variances V_h . To identify such cases, one simple diagnostic is the ratio $\{\hat{V}(\hat{V}_h)\}^{-1} \{\hat{\beta}_1^2 \hat{V}(\hat{V}_{wh}) + \hat{\sigma}_{qqh}\}$, i.e., the ratio of estimators of the variances of the approximate distributions of $\hat{V}_h - V_h$ and $\beta_1 \hat{V}_{wh} - V_h$, respectively, under model (3.1)–(3.2). If this ratio is substantially less than unity, then use of \hat{d}_{ws} may be indicated.

In addition, the performance of the estimator \hat{d}_{ws} depends heavily on the magnitude of $\hat{\sigma}_{uu}$ relative to the variability of the true within-PSU variances V_{wh} . Define an estimator of the reliability ratio (Fuller 1987, page 3)

$$\hat{\kappa}_{xx} = \max \left\{ 0, \left[\sum_{h=1}^L (\hat{V}_{wh} - \hat{V}_w)^2 \right]^{-1} \left[\sum_{h=1}^L (\hat{V}_{wh} - \hat{V}_w)^2 - \hat{\sigma}_{uu} \right] \right\}.$$

The values of $\hat{\kappa}_{xx}$ are between 0 and 1; and values of $\hat{\kappa}_{xx}$ close to unity indicate relatively small errors in the estimation of within-PSU variances. Conversely, small values of $\hat{\kappa}_{xx}$ (e.g., $\hat{\kappa}_{xx} < 0.7$) may indicate that the methods of Sections 3.1–3.2 may not perform well, due to the relatively large sampling errors in the auxiliary information \hat{V}_{wh} . The numerical work in Sections 4 and 5 below will consider these diagnostics further.

The work in this section is based on the assumption that $\sigma_{qq} > 0$. One may develop related diagnostics applicable to the case of no equation errors, i.e., $\sigma_{qq} = 0$; details are available from the authors.

4. A simulation study

4.1 Design of the study

We now use a simulation study to evaluate the properties of our degrees-of-freedom estimators, and related variates, under moderate-sample-size conditions. We set up the simulation procedure as follows.

We considered four sets of V_h values from the NHANES III example for the Mexican-American subpopulation introduced in Section 1.1. Those four sets of V_h are the estimated \hat{V}_h values from the variables BMPWT, HAR3, TCRESULT and HDRESULT, respectively, and are listed in Table 4.1. For each case, we used $(\beta_0, \beta_1) = (0, 1)$ and $\sigma_{qq} = 0$, in keeping with the results of Section 3, and thus $V_{wh} = V_h$. Then, for each $h = 1, \dots, L$, we obtained 10,000 realizations of the initial estimators $(\hat{Y}_{h1}, \hat{Y}_{h2}, \hat{V}_{wh1}, \hat{V}_{wh2})$ by assuming that the \hat{Y}_{hi} are distributed as a normal random variable with mean zero and variance $2^{-1} V_h$; that $V_{wh}^{-1}(m_{hi} - 1)\hat{V}_{whi}$ is distributed as a chi-square random variable with $m_{hi} - 1$ degrees of freedom, where $m_{hi} = 11$ for all h and i ; and the \hat{Y}_{hi} and \hat{V}_{whi} are mutually independent. Note that in our data from NHANES III, the average number of secondary units for each PSU i in stratum h is about 11. For each replication, we computed $\hat{V}_h = (\hat{Y}_{h1} - \hat{Y}_{h2})^2$ and $\hat{V}_{wh} = 2^{-1}(\hat{V}_{wh1} + \hat{V}_{wh2})$, and then carried out an errors-in-variables regression of \hat{V}_h on \hat{V}_{wh} with measurement error variance $\hat{\sigma}_{uuh} = \widehat{\text{Var}}(\hat{V}_{wh})$ using formula (2.2). This produced the coefficient estimators $(\hat{\beta}_0, \hat{\beta}_1)$, and the degrees-of-freedom estimators \hat{d}_{ms} and \hat{d}_{ws} .

Table 4.1
“True” variances V_h used in simulation studies

Stratum	Case 1	Case 2	Case 3	Case 4
1	0.00E+00	0.00E+00	0.00E+00	0.00E+00
2	0.00E+00	0.00E+00	0.00E+00	0.00E+00
3	1.56E-04	7.67E-05	1.45E-02	1.76E-02
4	2.01E-04	3.57E-06	5.60E-02	4.55E-03
5	2.82E-04	4.88E-07	1.54E-03	2.91E-03
6	4.36E-04	0.00E+00	3.73E-03	8.60E-04
7	7.30E-04	2.14E-06	1.69E-02	1.13E-05
8	8.80E-04	1.30E-05	2.72E-02	1.40E-03
9	1.65E-03	1.16E-06	9.24E-03	1.35E-04
10	1.70E-03	9.46E-07	2.24E-03	1.77E-03
11	2.73E-03	0.00E+00	2.54E-04	1.32E-03
12	2.91E-03	5.40E-06	2.75E-02	6.40E-03
13	4.95E-03	3.73E-07	1.15E-02	5.38E-03
14	7.25E-03	2.90E-04	3.75E-02	6.97E-02
15	9.06E-03	9.81E-05	3.46E-01	7.58E-01
16	1.14E-02	7.47E-06	1.54E-02	4.75E-03
17	2.69E-02	9.65E-05	7.99E-02	1.01E-03
18	4.00E-02	1.12E-04	1.44E-01	1.77E-01
19	4.27E-02	2.68E-06	8.59E-02	3.88E-02
20	6.05E-02	7.57E-06	2.68E+00	7.18E-02
21	6.45E-02	1.17E-04	1.65E-01	4.52E-04
22	1.08E-01	1.05E-04	5.41E-01	1.98E-03

4.2 Coverage rates of t -based confidence intervals

For the four specified cases, Table 4.2 presents the simulated non-coverage probabilities obtained for t -based confidence intervals for the population mean \bar{Y} that used the corresponding \hat{d} . For the severely heterogeneous cases (Cases 3 and 4), none of the degrees of freedom measures (not even the true d) leads to confidence intervals with coverage rates meeting the nominal rates $1 - \alpha$. That is, in extreme cases, the general Satterthwaite approach can be problematic for construction of confidence intervals, regardless of whether d , \hat{d}_{ms} , or \hat{d}_{ws} is used to determine the t multiplier.

For Cases 1 and 2, the V_h values display less severe heterogeneity than in Cases 3 and 4. Table 4.2 shows that the simulated coverage probabilities with the true d for these two cases are slightly above 0.95. This overcoverage may be attributable to the fact that the variance estimator $\hat{V}(\bar{Y})$ is not distributed exactly as a multiple of a χ^2_d random variable, due to the heterogeneity of the V_h . Use of the standard degrees-of-freedom term $n - L$ or the modified estimator \hat{d}_{ms} produces confidence intervals with coverage rates below the nominal level of 95%. On the other hand, use of our auxiliary-data-based term \hat{d}_{ws} gives simulation based coverage rates close to the nominal 0.95 level.

Tables 4.3a and 4.3b display the empirical distributions of \hat{d} and $2t_{\hat{d}}$ for the estimators \hat{d}_{ms} and \hat{d}_{ws} . The simulated

standard deviation of $t_{\hat{d}_{ws}}$ is smaller than that of $t_{\hat{d}_{ms}}$. In addition, the mean and median of $t_{\hat{d}_{ws}}$ are slightly larger than those of $t_{\hat{d}_{ms}}$. This is consistent with the undercoverage of the intervals based on $t_{\hat{d}_{ms}}$. Thus, under conditions similar to those for Cases 1 and 2 (or under conditions with less heterogeneity of V_h), it is worthwhile to consider the use of \hat{d}_{ws} as a degrees-of-freedom estimator.

5. Application to a health survey

5.1 Preliminary model checks

We applied our proposed methods to the NHANES III data described in Section 1. It is important to check the modeling assumptions before we apply the proposed stability measures. First, for the Mexican-American sub-population described in Section 1, Table 5.1 gives values of $\hat{\kappa}_{xx}$ for the four variables which all have $\hat{\kappa}_{xx}$ values greater than 0.7.

Second, Figure 5.1 displays the scatter plots of \hat{V}_h against \hat{V}_{wh} for the four variables with equal scales used for the horizontal and vertical axes. It shows that a linear relationship for the corresponding variables is plausible even if the relation would not be perfect and there are some outliers. Consequently, those four variables might be appropriate for the auxiliary-data-based method developed in Sections 2 and 3.

Table 4.2

Observed non-coverage rates for nominal 95% confidence intervals with $V_h = V_{wh}$ in simulation study

	Case 1	Case 2	Case 3	Case 4
True d_S	6.26	6.04	2.38	2.20
Non-Coverage with t_{d_i}	0.0428	0.0443	0.0162	0.0164
Non-Coverage with t_{n-L}	0.0744	0.0788	0.1220	0.1263
Non-Coverage with $t_{\hat{d}_{mS}}$	0.0552	0.0567	0.0911	0.0905
Non-Coverage with $t_{\hat{d}_{wS}}$	0.0428	0.0466	0.0224	0.0220

Table 4.3a

Means and quantiles of degrees-of-freedom estimators \hat{d}_{mS} and \hat{d}_{wS} : Cases 1 and 2

Cases	True d	Est.	¹ Mean \hat{d}	SD(\hat{d})	² Q(0.05)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.95)
1	6.26	\hat{d}_{mS}	9.33	3.33	4.45	6.86	9.01	11.41	15.30
		\hat{d}_{wS}	6.52	0.82	5.06	5.99	6.57	7.10	7.78
2	6.04	\hat{d}_{mS}	8.87	2.95	4.35	6.69	8.72	10.97	13.99
		\hat{d}_{wS}	6.34	0.96	4.67	5.69	6.42	7.06	7.80

¹ Mean denotes the average of the estimates, taken across all 10,000 replications.² Q(.) indicates the quantile of the estimator, taken across all 10,000 replications.

Table 4.3b

Simulated non-coverage probabilities; and means and quantiles of t -multipliers for nominal 95% confidence intervals: Unequal true variances, cases 1 and 2

Cases	Est.	¹ $1 - \hat{\alpha}$	² M($2t_{\alpha}$)	SD($2t_{\alpha}$)	³ Q(0.05)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.95)
1	\hat{d}_{mS}	0.0552	4.62	0.36	4.26	4.38	4.52	4.75	5.37
	\hat{d}_{wS}	0.0428	4.83	0.16	4.64	4.72	4.80	4.90	5.13
	$n - L$	0.0744	4.15						
	True d_S	0.0428	4.85						
2	\hat{d}_{mS}	0.0567	4.66	0.36	4.29	4.41	4.55	4.78	5.41
	\hat{d}_{wS}	0.0466	4.87	0.21	4.64	4.72	4.83	4.97	5.28
	$n - L$	0.0788	4.15						
	True d_S	0.0443	4.89						

¹ $1 - \hat{\alpha}$ is the simulated non-coverage probability of confidence intervals computed using estimated d.f.'s² M($2t_{0.975}$) is the average of twice of the 97.5% t -percentile value³ Q(.) indicates the quantile of $2t_{0.975, \hat{d}}$, taken across all replications.

Table 5.1

 $\hat{\kappa}_{xx}$, estimates of model parameters, model diagnostics, and degrees of freedom estimates for four NHANES III variables (Mexican-American (Age 20-29) subgroup)

Variables	$\hat{\kappa}_{xx}$	$\tilde{\beta}_0$	se($\tilde{\beta}_0$)	$\tilde{\beta}_1$	se($\tilde{\beta}_1$)	Simulation based p-value for $H_0: \beta_0 = 0$	Simulation based p-value for $H_0: \beta_1 = 1$	$\hat{\sigma}_{qq}$	\hat{r}_{qq}	\hat{d}_{mS}	\hat{d}_{wS}
BMPWT	0.75	-0.0013	0.0039	1.135	0.5429	0.3815	0.3541	-0.000	-0.43	15.49	10.04
HAR3	0.75	-0.000009	0.000012	1.095	0.3991	0.4229	0.3400	0.000	-0.83	14.94	8.30
TCRESULT	0.88	-0.146	0.0493	2.879	0.6252	0.0606	0.2259	-0.178	-0.77	5.88	6.59
HDRESULT	0.90	-0.042	0.0098	6.650	0.9988	<0.0001	0.1506	-0.017	-0.91	5.45	5.93

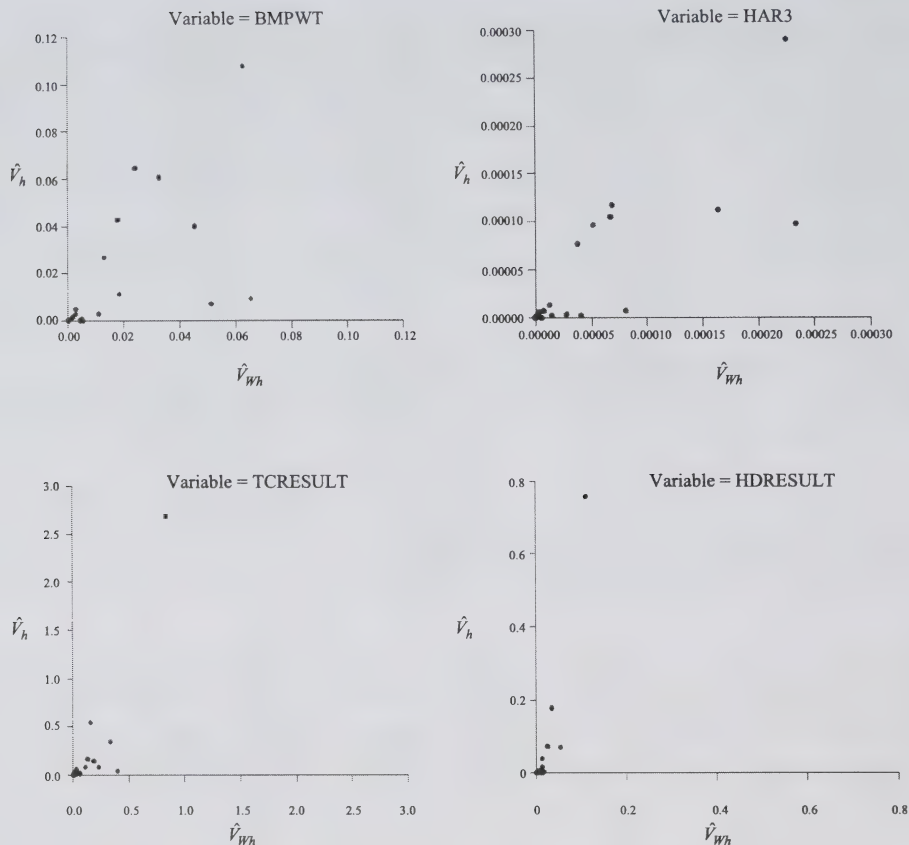


Figure 5.1 Plot of \hat{V}_{Wh} vs. \hat{V}_h for M-A (Age 20-29), Variable = BMPWT

5.2 An *ad hoc* test of $\bar{\sigma}_{qq} = 0$ under condition (C.1)

For all four variables considered in Table 5.1, the direct estimates $\hat{\sigma}_{qq}$ of equation error variance (3.4) were negative or close to zero. That suggests that our χ^2 -based estimator of σ_{eeh} as given in Section 3.1 might be too conservative or that $\bar{\sigma}_{qq}$ is indeed close to zero. This suggests that we need to re-examine the distributional assumption (C.1) in the NHANES III example. To do this, we considered the simulated distribution of $\hat{r}_{qq} \stackrel{\text{def}}{=} \hat{\sigma}_{qq} / \hat{\sigma}_{ee}$, where division by $\hat{\sigma}_{ee}$ is used to avoid scale problems. The conditions and simulation design were as described in Section 4.1.

Table 5.2 reports results for $\hat{\sigma}_{ee}$ from expression (3.5), and $\hat{\sigma}_{qq}$ computed from expression (3.4) with $\hat{\beta}_0$ set equal to zero and with $\hat{\beta}_1$, computed from expression (3.3). Table 5.2 reports the mean, standard deviation and selected quantiles of the simulated distribution of \hat{r}_{qq} for the four variables. Table 5.3 reports the corresponding quantities for \hat{r}_{qq} , computed from $\hat{\sigma}_{qq}$ given by expression (3.4) and with $\hat{\beta}_0$ and $\hat{\beta}_1$ computed from expression (3.3).

The results reported in Tables 5.2 and 5.3 lead to an *ad hoc* test of $H_0: \sigma_{qq} = 0$. Specifically, if the observed ratio \hat{r}_{qq} falls above the upper 0.95 simulated quantile, then the assumption that $\bar{\sigma}_{qq} = 0$ may be problematic. Conversely, an observed \hat{r}_{qq} below the .05 simulated quantiles in Tables 5.2 or 5.3 might indicate that $\hat{\sigma}_{eeh}$ is conservative, or may indicate violation of other parts of condition (C.1).

From Table 5.1, the values of \hat{r}_{qq} for the variables are between -0.91 to -0.43. Except for HDRESULT, we do not have any strong evidence of violation of the model assumptions. However, for HDRESULT, the ratio $\hat{r}_{qq} = -0.91$ falls between the 0.01 and 0.05 quantiles reported in Table 5.2 and 5.3 for case 4. In general, values of \hat{r}_{qq} that fall above the 0.95 or 0.99 quantiles of Tables 5.2 or 5.3 would be consistent with values of $\bar{\sigma}_{qq}$ greater than zero. The observed value $\hat{r}_{qq} = -0.91$ is not necessarily consistent with $\bar{\sigma}_{qq} > 0$, but may indicate violation of one or more conditions in (C.1)-(C.4).

Table 5.2

Means and quantiles of $\hat{r}_{qq} = \hat{\sigma}_{ee}^{-1} \hat{\sigma}_{qq}$, ($\beta_0 = 0$)

Cases	$^1M(\hat{r}_{qq})$	$SD(\hat{r}_{qq})$	$^2Q(0.01)$	$Q(0.05)$	$Q(0.10)$	$Q(0.25)$	$Q(0.50)$	$Q(0.75)$	$Q(0.90)$	$Q(0.95)$	$Q(0.99)$
1	-0.50	0.66	-1.71	-1.30	-1.15	-0.99	-0.79	0.16	0.54	0.60	0.65
2	-0.48	0.68	-1.72	-1.32	-1.16	-0.99	-0.76	0.23	0.57	0.62	0.66
3	-0.19	0.42	-1.01	-0.84	-0.74	-0.53	-0.20	0.17	0.38	0.46	0.55
4	-0.20	0.39	-1.00	-0.82	-0.72	-0.51	-0.20	0.11	0.34	0.44	0.56

¹ M denotes the average of the estimates, taken across all 10,000 replications.² Q(.) indicates the quantile of the estimator, taken across all 10,000 replications.

Table 5.3

Means and quantiles of $\hat{r}_{qq} = \hat{\sigma}_{ee}^{-1} \hat{\sigma}_{qq}$.

Cases	$^1M(\hat{r}_{qq})$	$SD(\hat{r}_{qq})$	$^2Q(0.01)$	$Q(0.05)$	$Q(0.10)$	$Q(0.25)$	$Q(0.50)$	$Q(0.75)$	$Q(0.90)$	$Q(0.95)$	$Q(0.99)$
1	-0.56	0.62	-1.85	-1.34	-1.17	-1.00	-0.80	0.05	0.38	0.44	0.52
2	-0.56	0.62	-1.91	-1.37	-1.18	-1.00	-0.78	0.06	0.35	0.42	0.50
3	-0.24	0.42	-1.16	-0.90	-0.79	-0.57	-0.22	0.12	0.29	0.36	0.45
4	-0.24	0.38	-1.09	-0.87	-0.75	-0.53	-0.22	0.06	0.25	0.33	0.44

¹ M denotes the average of the estimates, taken across all 10,000 replications.² Q(.) indicates the quantile of the estimator, taken across all 10,000 replications.

5.3 Coefficient estimates and degrees-of-freedom estimates

Because our data were consistent with $\bar{\sigma}_{qq} = 0$ for all four cases, we used the methods of Fuller (1987, page 124) to produce estimates of β_0 and β_1 appropriate for a model (3.1)–(3.2) with no equation error; details are available from the authors. Table 5.1 also reports the resulting coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, and their standard errors, $se(\hat{\beta}_0)$ and $se(\hat{\beta}_1)$. Recall from Section 3.1 that under model (3.1)–(3.2), if $\beta_0 = 0$ and $\beta_1 \neq 0$, then each stratum variance V_h is a constant multiple of the within-PSU variance V_{wh} , and $\hat{\mu}_{WS}$ in (1.8) may be an appropriate estimator of d . Section 5.2 already considered the condition $\bar{\sigma}_{qq} = 0$. To test the null hypothesis $H_0: \beta_0 = 0$, we use the test statistic, $t_0 = \hat{\beta}_0/se(\hat{\beta}_0)$. In some practical errors-in-variables work, quantities like t_0 are compared with a standard normal or t reference distribution. However, simulation work based on the four cases from Section 4.1 indicated that the null distribution of t_0 deviated substantially from these customary reference distributions. This is due to the very skewed distributions of the response variables \hat{V}_h used in the errors-in-variables regression. Consequently, we used standard methods to develop a simulation-based reference distribution for t_0 . Column 7 of Table 5.1 reports the resulting left-tailed p -value. (Due to

negative point estimates $\hat{\beta}_0$, we have chosen to report the left-tailed p -values here. In other cases, it may be of interest to report right-tailed or two-tailed p -values for β_0). There is strong evidence against $H_0: \beta_0 = 0$ for the variable HDRESULT, and the moderate evidence against $H_0: \beta_0 = 0$ for TCRESULT. Thus, it may not be appropriate to use $\hat{\mu}_{WS}$ for these two variables. Now consider the slope coefficient β_1 , and suppose that $\sigma_{qqh} = 0$ so $q_h = 0$ with probability one. Then expressions (1.5) and (3.1), and the nonnegativity of V_{bh} implies that $0 \leq V_{bh} = V_h - V_{wh} = \beta_0 + (\beta_1 - 1)V_{wh}$. Consequently, if $\beta_0 = 0$, then $\beta_1 \geq 1$ and $\beta_1 = 1$ is equivalent to $V_h = V_{wh}$. This final condition is of practical interest because some authors have noted cases in which V_{bh} is small relative to V_{wh} , or equivalently, $V_h \doteq V_{wh}$. See for example, Wolter (1985, page 46). To test $H_0: \beta_1 = 1$ against the one-sided alternative $H_1: \beta_1 > 1$, we used the statistic $t_1 = (\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$. For reasons similar to those for t_0 , we developed simulation-based reference distributions for t_1 under each of Cases 1 through 4. Column 8 of Table 5.1 reports the resulting one-tailed p -values.

The last two columns of Table 5.1 report the degree-of-freedom estimators \hat{d}_{MS} and \hat{d}_{WS} . For HAR3 and BMPWT, \hat{d}_{MS} gives substantially larger values than \hat{d}_{WS} .

6. Discussion

This paper has considered estimation of a degrees-of-freedom term d used to quantify the variability of a standard design based variance estimator $\hat{V}(\hat{Y})$. The fundamental issue is that under a design involving heterogeneous stratum-level variances and small numbers of primary sample units selected per stratum, the Satterthwaite-type estimator \hat{d}_{mS} may perform poorly. We developed an alternative estimator \hat{d}_{wS} based on within-primary-sample unit variance estimators \hat{V}_{wh} . This alternative estimator is a solution to an unbiased estimating equation (1.1) for d , provided the proportionality condition (1.6) is satisfied. Also, the variance of the approximate distribution of \hat{d}_{wS} is smaller than that of \hat{d}_{mS} , provided the number of secondary sample units selected within each primary unit is large, in the sense defined by Result 2.2.

Section 3 developed errors-in-variables methods for testing the adequacy of the proportionality condition (1.6), and suggested some related diagnostics. The simulation study in Section 4, in conjunction with the data analysis in Section 5, indicated that under moderate amounts of heterogeneity, \hat{d}_{wS} can perform better than \hat{d}_{mS} , in terms of the distributional properties of these estimators of d , and in terms of the coverage rates and widths of associated confidence intervals for the population totals Y . However, as one would expect from standard large-sample theory, neither estimator performs well under severe heterogeneity.

One could in principle consider use of the errors-in-variables estimators $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_{qq})$, in conjunction with the \hat{V}_h and \hat{V}_{wh} , to construct an alternative estimator of d that will be consistent under the general errors-in-variables model (3.1)-(3.2), and will not require the restrictive condition (1.6). However, simulation results in Jang (1996) indicated that the resulting estimator \hat{d}_{EIV} , say, did not perform well under the design conditions used in Section 5.

The principal results of Sections 1 through 3 extend readily from the within-primary-unit variances V_{wh} to more general auxiliary variables X_h . For such extensions, the principal issues remain the adequacy of the proportionality approximation (1.6); and the amount of sampling error in the auxiliary estimators \hat{X}_h , say, relative to the error in the basic stratum-level variance estimator \hat{V}_h .

Acknowledgements

The authors thank the U.S. National Center for Health Statistics for providing access to the NHANES III dataset, and thank V.L. Parsons, C. Johnson and L.R. Curtin for sharing a wealth of information regarding the NHANES III. This research was supported in part by the U.S. National Center for Health Statistics. The views expressed in this

paper are those of the authors and do not necessarily represent the policies of the U.S. National Center for Health Statistics or the U.S. Bureau of Labor Statistics.

Appendix A

Proof of result 2.2

Consider a nonlinear function $B^{-1}A^2$ of two estimators A and B with means μ_A and μ_B , respectively. Then, the variance of the leading term of a Taylor expansion of $B^{-1}A^2$ is

$$\frac{4\mu_A^2}{\mu_B^2} \text{Var}(A) - 4\frac{\mu_A^3}{\mu_B^3} \text{Cov}(A, B) + \frac{\mu_A^4}{\mu_B^4} \text{Var}(B). \quad (\text{A.1})$$

Now we define the following two estimators: $L^{-1}\hat{d}_{S1} = B_1^{-1}A_1^2$ and $L^{-1}\hat{d}_{S2} = B_2^{-1}A_2^2$, where $A_1 = L^{-1}\sum_{h=1}^L \hat{V}_h$, $B_1 = L^{-1}\sum_{h=1}^L \hat{V}_h^2$, $A_2 = L^{-1}\sum_{h=1}^L \hat{V}_{wh}$, and $B_2 = L^{-1}\sum_{h=1}^L \{\hat{V}_{wh}^2 - \widehat{\text{Var}}(\hat{V}_{wh})\}$.

Assume conditions (C.1), (C.2) and (C.3). In addition, define $\hat{F}_{L\hat{d}_{S1}}$ and $\hat{F}_{L\hat{d}_{S2}}$ to be the leading terms of Taylor expansions of $L^{-1}\hat{d}_{S1} - \mu_{B_1}^{-1}\mu_{A_1}^2$ and $L^{-1}\hat{d}_{S2} - \mu_{B_2}^{-1}\mu_{A_2}^2$, respectively. Also, recall that if D is distributed as a chi-square random variable on d degrees of freedom, then $V(D) = 2d$, $E(D^3) = d(d+2)(d+4)$, and $V(D^2) = 8d(d+2)(d+3)$. Then the corresponding components of $\text{Var}(\hat{F}_{L\hat{d}_{S1}})$ and $\text{Var}(\hat{F}_{L\hat{d}_{S2}})$ in (A.1) are

$$\text{Var}(A_1) = 2L^{-2} \sum_{h=1}^L V_h^2,$$

$$\text{Var}(A_2) = 2(m_0 - 1)^{-1} L^{-2} \sum_{h=1}^L V_{wh}^2$$

$$\text{Var}(B_1) = 96L^{-2} \sum_{h=1}^L V_h^4,$$

$$\text{Var}(B_2) = 8(m_0 - 1)^{-2} (m_0 + 1) L^{-2} \sum_{h=1}^L V_{wh}^4$$

$$\text{Cov}(A_1, B_1) = 12L^{-2} \sum_{h=1}^L V_h^3,$$

and

$$\text{Cov}(A_2, B_2) = 4(m_0 - 1)^{-1} L^{-2} \sum_{h=1}^L V_{wh}^3. \quad (\text{A.2})$$

Since we assume $n_h = 2$ and $m_h = m_0$ for all $h = 1, 2, \dots, L$, we have

$$L^{-1}\hat{d}_{mS} = L^{-1}(3L+14)^{-1}(9L)\hat{d}_{S1} \quad (\text{A.3})$$

and

$$L^{-1}\hat{d}_{wS} = L^{-1}\hat{d}_{S2}. \quad (\text{A.4})$$

Under condition (1.6), $\mu_{A1} = \beta_1 \mu_{A2}$,

$$\mu_{B1} = 3\beta_1^2 \mu_{B2},$$

$$\text{Var}(A_1) = (m_0 - 1)\beta_1^2 \text{Var}(A_2),$$

$$\text{Var}(B_1) = 12(m_0 + 1)^{-1}(m_0 - 1)^2 \beta_1^4 \text{Var}(B_2)$$

and

$$\text{Cov}(A_1, B_1) = 3(m_0 - 1)\beta_1^3 \text{Cov}(A_2, B_2) \quad (\text{A.5})$$

Substituting (A.5) into (A.1) leads to,

$$\begin{aligned} \text{Var}(\hat{F}_{L\hat{d}_{S1}}) &= \frac{4}{9}(m_0 - 1) \frac{\mu_{A_2}^2}{\mu_{B_2}^2} \text{Var}(A_2) \\ &\quad - \frac{4}{9} \frac{\mu_{A_2}^3}{\mu_{B_2}^3} (m_0 - 1) \text{Cov}(A_2, B_2) \\ &\quad + \frac{4(m_0 - 1)^2}{27(m_0 + 1)} \frac{\mu_{A_2}^4}{\mu_{B_2}^4} \text{Var}(B_2) \\ &= \frac{1}{9}(m_0 - 1)a - \frac{1}{9}(m_0 - 1)b + \frac{4(m_0 - 1)^2}{27(m_0 + 2)}c \quad (\text{A.6}) \end{aligned}$$

where $\text{Var}(L^{-1}\hat{d}_{WS}) = a - b + c$. With large L , $\text{Var}(\hat{F}_{L\hat{d}_{S1}}) = (m_0 - 1)a - (m_0 - 1)b + \{3(m_0 + 2)\}^{-1}4(m_0 - 1)^2c$. Thus for large L , $V(\hat{F}_{L\hat{d}_{S1}}) - V(\hat{F}_{L\hat{d}_{WS}}) \doteq (m_0 - 2)a - (m_0 - 2)b + \{3(m_0 + 2)\}^{-1}(4m_0^2 - 11m_0 - 2)c$. Therefore, $\lim_{L \rightarrow \infty} V_{Lm} - \lim_{L \rightarrow \infty} V_{LW} \geq 0$ if $m_0 \geq \lim_{L \rightarrow \infty} \{2(3a - 3b + 4c)\}^{-1}\{11c + \sqrt{144a^2 + 144b^2 + 153c^2 - 288ab + 216ac - 216bc}\}$. In particular, $\lim_{L \rightarrow \infty} V_{Lm} - \lim_{L \rightarrow \infty} V_{LW}$ becomes greater than or equal to zero when $m_0 = 10$ regardless of values of a , b , and c . Because it is an increasing function in m_0 , for all values of $m_0 \geq 10$, $\lim_{L \rightarrow \infty} V_{Lm} \geq \lim_{L \rightarrow \infty} V_{LW}$.

References

- Casady, R., Dorfman, A.H. and Wang, S. (1998). Confidence intervals for sub-domain parameters when the sub-domain sample size is random. *Survey Methodology*, 24, 57-67.
- Cochran, G.C. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.
- Eltinge, J.L., and Jang, D. (1996). Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology*, 22, 157-165.
- Fuller, W.A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.
- Isaki, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78, 117-123.
- Jang, D. (1996). *Stability of Variance Estimators Under Complex Sampling Designs*. Unpublished Ph.D. dissertation, Department of Statistics, Texas A&M University, College Station, Texas.
- Kendall, M., Stuart, A. and Ord, J.K. (1983). *The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time Series*. New York: Macmillan.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.*, 9, 1010-1019.
- National Center for Health Statistics (1996). NHANES III Reference Manuals and Reports, CD-ROM GPO, 017-022-1358-4. Washington, D.C.: United States Government Printing Office.
- Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.
- Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.
- Smith, H.F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, 9, 211-212.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Semiparametric regression model for complex survey data

Zilin Wang and David R. Bellhouse¹

Abstract

A semiparametric regression model is developed for complex surveys. In this model, the explanatory variables are represented separately as a nonparametric part and a parametric linear part. The estimation techniques combine nonparametric local polynomial regression estimation and least squares estimation. Asymptotic results such as consistency and normality of the estimators of regression coefficients and the regression functions have also been developed. Success of the performance of the methods and the properties of estimates have been shown by simulation and empirical examples with the Ontario Health Survey 1990.

Key Words: Complex survey; Domain estimates; Nonparametric regression; Smoothing.

1. Introduction

In practice, many surveys are used to explore a relationship between a response variable and explanatory variables and to build predictive models. Hence, it is necessary to develop techniques that apply stochastic regression models to survey data. Although nonparametric regression techniques have been widely applied in many fields of statistics, not much attention has been paid to them in the field of complex surveys due to the complexity of the data structure. The correlation induced by clustering and unequal probabilities of selection of the sample cause survey data to be neither independent nor identically distributed. As a result, standard nonparametric regression methods are often inappropriate for analyzing sample survey data.

There is some work, for instance Breidt and Opsomer (2000), Montanari and Ranalli (2005), and Zheng and Little (2004), on nonparametric regression techniques that have been developed for survey data. However, as in the conventional way of applying regression techniques, most of this work uses model-assisted approaches to estimate descriptive population quantities and parameters related to the descriptive quantities. In this paper, we are interested in the application of nonparametric regression techniques to exploring the relationship between the response variable and covariates, as well as prediction using auxiliary information. Bellhouse and Stafford (2001) extended a local polynomial regression technique to conduct flexible regression modelling for complex survey data. However, their paper dealt only with a simple nonparametric regression function. Here we extend their enquiry to a case of several independent variables, including indicator variables that often appear in regression analysis for survey data.

We consider a partially linear semi-parametric regression function defined as $E(y | \mathbf{X}, \mathbf{z}) = \mathbf{X}\beta + G(\mathbf{z})$, where $G(\cdot)$ is an arbitrary function and β is an unknown

p -dimensional parameter vector. In this semi-parametric regression model, the explanatory variables are represented separately in two parts: a nonparametric part and a parametric linear part. It is of interest to estimate both the functional form of the nonparametric part of the model and the parameters that are included in the parametric part of the model. We put the categorical explanatory variables and continuous variables with assumed linear dependence in the parametric part of the model, $\mathbf{X}\beta$, and a variable with little information on the functional form in the nonparametric part of the model, $G(\mathbf{z})$. This partial linear semi-parametric model not only has a priori motivation as a data analytic tool and retains an important interpretive feature, it also eases the high dimensional problem created by factors and some covariates by including them in the parametric part of the model.

A similar model has been developed for independently and identically distributed data independently by Robinson (1988) and Speckman (1988). In these papers, the estimation is conducted in three steps. In the first step, the means of the response variable and the parametric independent variables, conditional on the nonparametric variable, are treated as a function of that variable and smoothed; in the second step, the linear coefficients are estimated by regressing the residuals from the smoothed response variable on the residuals from the smoothed parametric covariates; finally, the difference between the response variable and its prediction from the regression model is smoothed in a similar manner to provide an estimate of the nonparametric part of the regression function. It has been shown in Robinson (1988) and Speckman (1988) that the resulting estimators are root- n consistent when the model is correct and the data points are independent and identically distributed. The objective of our paper is to apply this smoothing technique to survey data while allowing for a complex sampling scheme.

1. Zilin Wang, Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada, N2L 3C5. E-mail: zwang@wlu.ca; David R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, Canada, N6A 5B7. E-mail: bellhouse@stats.uwo.ca.

We use the local polynomial regression estimation technique developed in Bellhouse and Stafford (2001) to conduct all the smoothing during the estimation process. A key element in accomplishing the local polynomial regression technique from Bellhouse and Stafford (2001) is binning, which follows the work of Bellhouse and Stafford (1999) in density estimation. In many survey data sets, a continuous variable may be naturally binned; for example age may be recorded as age last birthday. In general, bins correspond to the disjoint sets of values of a continuous covariate, and thus can be regarded as domains. At the level of the sample, we estimate the domain mean of the variable of interest by dividing the weighted sum of the variable within the domain by the sum of the weights within the domain. In Bellhouse and Stafford (2001), the response variable is binned according to the values of the covariate, and discretized, and the domain means of the response variable are smoothed to obtain the regression function. When the sample size is large and the number of bins is relatively small, then estimators based on binning are functions of domain estimators whose inferential properties can be readily derived from results in Shao (1996) and Serfling (1980). One of the practical advantages to binning is that it can reveal information on an obscured trend in a complex survey, which is sometimes quite important when the scale of the complex survey data set is large. There are, usually, multiple observations at each set of covariate values in these data sets.

An example that illustrates these features of binned data is taken from the Ontario Health Survey. The survey was conducted by Statistics Canada in 1990 with 61,239 individuals living in Ontario, Canada. The data were obtained by a stratified two-stage clustered design. The strata were the urban and rural areas covered by each of the

public health units in the province of Ontario. Within each stratum enumeration areas were randomly selected, as were households within each enumeration area. The purpose of this survey is to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of mortality and morbidity in Ontario. In this example, we examine people's weight as a function of age. In the Ontario Health Survey, age was given only to age last birthday. The measurement we use for a proxy of weight here is called body mass index (BMI) which is calculated as weight in kilograms divided by the square of height in meters. BMI is used as one of the indicator of a person's obesity level. Normally, a person with a BMI below 18 is considered underweight and a BMI greater than 30 suggests obesity. BMI is used as an appropriate measure only for all persons between the ages of 18 and 64 with the exception of pregnant and breast feeding women. Consequently the sample size is reduced to 44,457 eligible respondents that have 47 distinct possible ages or bins.

In the left panel of Figure 1, the age trend of body mass index is plotted. It is readily seen that the "black cloud"-like scatterplot masks the relationship between age and body mass index. Now, if we calculate mean of the body mass index at each distinct point of age, and plot the binned mean estimates of the body mass index versus age, we can obtain the plot in the right panel of Figure 1. It is obvious that a binned mean provides more visual information than the raw data does. Large-scale data sets not only can result in non-informative plots, they also make the estimation process computationally very cumbersome. Hence, it is natural in complex survey data analysis to bin the data into domains according to distinct values of a discretized covariate. Further, estimators from binning are functions of domain estimators.

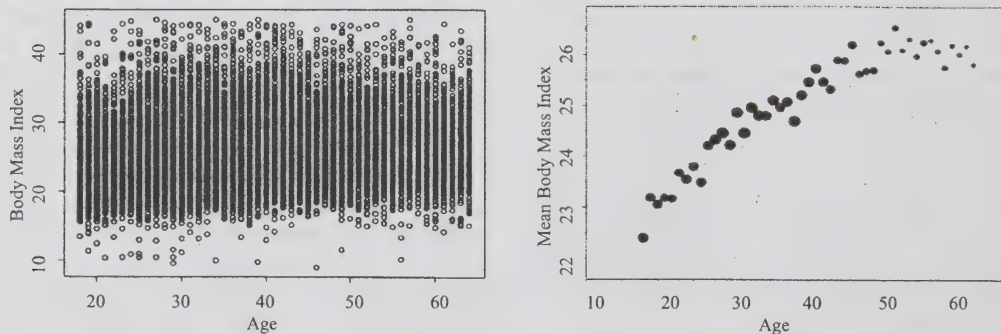


Figure 1 Comparison of the scatter plots of the binned and unbinned data from the Ontario health survey

One drawback to binning is that number of bins cannot grow asymptotically with the population if the data are naturally binned, as with the age variable in the above example. In such a case, the population level nonparametric estimators will remain biased as estimators of superpopulation functions due to a fixed bin size. In our framework, we assume that the bins induced by the distinct values of the covariate are the same in the population as in the sample; similarly in smoothing we will take the bandwidth to be the same at the population level as at the sampling level. We will show that the sample estimators are design consistent estimators of the corresponding finite population parameters and functions, though not of their superpopulation counterparts. In the Ontario Health Survey data example, the same set of distinct ages appears in both the population and the sample.

The paper is organized as follows. Superpopulation working models leading to the estimation procedures in survey data are introduced in Section 2. In Section 3, we derive all the moments of the estimates obtained and establish some asymptotic results. A simulation study and an empirical illustration of the estimation method carried out using the 1990 Ontario Health Survey (1992) appear in Section 4 and Section 5. Section 6 concludes with a discussion of assumptions made and some future work. The Proofs of all lemmas and theorems in Section 3 are given in an appendix.

2. Semiparametric regression model and its estimation

We take a typical approach to complex survey data analysis. First, we assume a working model on the finite population under the assumption of independent observations. Model parameter estimates then become the finite population parameters, or census parameters, to be estimated from the survey sample. Once the finite population target parameters have been defined we assume a more realistic model on the finite population in order to obtain inferences about these parameters. This is done in the next section. Consider a finite population of size N with a vector of measurements (y_k, \mathbf{x}_k, z_k) attached to unit k , $k = 1, \dots, N$, where y_k represents an observation of the response variable and (\mathbf{x}_k, z_k) represents a vector of observations of the explanatory variables with length $p + 1$. As a working model we imagine that the response variable is generated by the following partial linear regression model,

$$\mathbf{Y} = G(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where \mathbf{Y} is the vector of responses and $\boldsymbol{\varepsilon}$ has entries that are independent and identically distributed with mean zero

and constant variance. The function $G(\cdot)$ is an arbitrary function of \mathbf{z} and $\boldsymbol{\beta}$ is an unknown p -dimensional parameter vector. The $N \times p$ matrix \mathbf{X} corresponds to the linear part of the model and contains either continuous or discrete explanatory variables which are random. The term $G(\mathbf{z})$ is the nonparametric part of the model. We assume that \mathbf{z} is non-stochastic and measured on a continuous scale, discretized into D distinct values. Additionally, it is imagined that $E(\boldsymbol{\varepsilon} | \mathbf{z}, \mathbf{X}) = \mathbf{0}$. There is no interaction between \mathbf{X} and \mathbf{z} in the model.

We are interested in estimating population level versions of $G(\cdot)$ and the parameters $\boldsymbol{\beta}$. We first develop expressions for these, guided by the estimation procedures in Robinson (1988) and Speckman (1988). In particular, we begin by taking the expectation of both sides of (1) conditional on \mathbf{z} :

$$E(\mathbf{Y} | \mathbf{z}) = E(\mathbf{X} | \mathbf{z})\boldsymbol{\beta} + G(\mathbf{z}). \quad (2)$$

Then we subtract (2) from (1) to obtain

$$\mathbf{Y} - E(\mathbf{Y} | \mathbf{z}) = (\mathbf{X} - E(\mathbf{X} | \mathbf{z}))\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (3)$$

To define the population version of $\boldsymbol{\beta}$ in (3), we will replace $E(\mathbf{Y} | \mathbf{z})$ and $E(\mathbf{X} | \mathbf{z})$ in (3) by their population level estimates and estimate $\boldsymbol{\beta}$ by the method of least squares.

For the population level estimates of $E(\mathbf{Y} | \mathbf{z})$ and $E(\mathbf{X} | \mathbf{z})$, we adopt the local polynomial smoother in Jones (1989), in which binning is an essential part of the operation. Let the discretized Z variable take values z_1, \dots, z_D ; let the vectors of means in the bins of z_1, \dots, z_D be $\bar{\mathbf{Y}} = (\bar{Y}_1, \dots, \bar{Y}_D)$ and $\bar{\mathbf{X}}_j = (\bar{X}_{j1}, \dots, \bar{X}_{jD})$ for $j = 1, \dots, p$, respectively. Also, let P_d be the population proportion of observations in the d^{th} bin for $d = 1, \dots, D$. Then denote the population smoothed conditional expectations of \mathbf{Y} and \mathbf{X}_j at the point z_d by $m_y(z_d)$ and $m_j(z_d)$, respectively. Given that $K(\cdot)$ is a kernel function satisfying $\int K(t) dt = 1$ and $\int K(t)^2 dt < \infty$ and h is the bandwidth and using the principle of local polynomial regression technique, we minimize

$$\sum_{d=1}^D \frac{P_d}{h} \{ \bar{Y}_d - \alpha_0 - \alpha_1(z'_d - z_d), \dots, -\alpha_q(z'_d - z_d)^q \}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \quad (4)$$

and

$$\sum_{d=1}^D \frac{P_d}{h} \{ \bar{X}_{jd} - \gamma_0 - \gamma_1(z'_d - z_d), \dots, -\gamma_q(z'_d - z_d)^q \}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \quad (5)$$

with respect to α 's and γ 's so that the population estimated (smoothed) conditional expectations of y and X_j

on z_d , $m_y(z_d)$ and $m_j(z_d)$, are the solutions of α_0 and γ_0 for equations (4) and (5). Specifically,

$$m_j(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_w \bar{\mathbf{X}}_j$$

and

$$m_y(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_w \bar{\mathbf{Y}}$$

where q is the degree of the polynomial smoother, \mathbf{e} is a $(q+1) \times 1$ vector in the form of $(1, 0, 0, \dots, 0)^T$, and \mathbf{Z} and \mathbf{K}_w are respectively defined as

$$\mathbf{Z} = \begin{pmatrix} 1 & z_1 - z_d & \dots & (z_1 - z_d)^q \\ 1 & z_2 - z_d & \dots & (z_2 - z_d)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_D - z_d & \dots & (z_D - z_d)^q \end{pmatrix} \quad (6)$$

and $\mathbf{K}_w = \text{diag}(\hat{P}_1 K((z_1 - z_d)/h), \dots, \hat{P}_D K((z_D - z_d)/h))/h$.

With the census estimators of the conditional expectations $m_j(z_d)$ and $m_y(z_d)$, we define a $N \times p$ matrix \mathbf{M}_x and a $N \times 1$ vector \mathbf{M}_y as,

$$\mathbf{M}_x = \begin{pmatrix} m_1(z_1) & m_2(z_1) & \dots & m_p(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_1) & m_2(z_1) & \dots & m_p(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_D) & m_2(z_D) & \dots & m_p(z_D) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_D) & m_2(z_D) & \dots & m_p(z_D) \end{pmatrix} \quad (7)$$

and

$$\mathbf{M}_y = \begin{pmatrix} m_y(z_1) \\ \vdots \\ m_y(z_1) \\ \vdots \\ m_y(z_D) \\ \vdots \\ m_y(z_D) \end{pmatrix}.$$

Note that the d^{th} blocks of \mathbf{M}_x and \mathbf{M}_y are of the dimensions of $N_d \times p$ and $N_d \times 1$, respectively, where N_d be number of observations that fall in the d^{th} bin and $\sum N_d = N$. Replacing the conditional expectation matrix, $E(\mathbf{X} | \mathbf{z})$, and vector, $E(\mathbf{Y} | \mathbf{z})$, in (3) with their estimates, \mathbf{M}_x and \mathbf{M}_y , and using the general estimating equations

framework suggested by Godambe and Thompson (1986) for the least squares estimation, we can obtain the finite population versions parameters (census estimators) of β , namely \mathbf{B} , by solving

$$\begin{aligned} \mathbf{u}(\theta) &= \sum_{k=1}^N (\mathbf{x}_k - \mathbf{M}_{xk})^T (y_k - M_{yk}) \\ &\quad - \sum_{k=1}^N (\mathbf{x}_k - \mathbf{M}_{xk})^T (\mathbf{x}_k - \mathbf{M}_{xk}) \mathbf{B} \\ &= \mathbf{0}_{p \times 1}, \end{aligned} \quad (8)$$

where \mathbf{M}_{xk} is the k^{th} row of the $N \times p$ matrix \mathbf{M}_x and M_{yk} is the k^{th} element of the $N \times 1$ vector \mathbf{M}_y . The finite population parameter vector θ^T is composed of $(\mathbf{B}^T, \mathbf{m}_x(\mathbf{z}), \mathbf{m}_y(\mathbf{z})^T)$, where $\mathbf{m}_x(\mathbf{z})$ is a vector of the form $(m_1(\mathbf{z})^T, \dots, m_p(\mathbf{z})^T)$ with $\mathbf{m}_j(\mathbf{z}) = (m_j(z_1), \dots, m_j(z_D))$ for $j = 1, \dots, p$ and $\mathbf{m}_y(\mathbf{z}) = (m_y(z_1), \dots, m_y(z_D))$. Hence, the closed form expression for the estimator (census parameter) \mathbf{B} is

$$\mathbf{B} = ((\mathbf{X} - \mathbf{M}_x)^T (\mathbf{X} - \mathbf{M}_x))^{-1} (\mathbf{X} - \mathbf{M}_x)^T (\mathbf{Y} - \mathbf{M}_y).$$

Once \mathbf{B} is obtained, the difference between the response variable \mathbf{Y} and the product $\mathbf{X}\mathbf{B}$ is treated as the dependent random variable and the function $G(\cdot)$ is estimated in accordance with the following model

$$\mathbf{Y} - \mathbf{X}\mathbf{B} = G(\mathbf{z}) + \varepsilon.$$

The finite population version of $G(\mathbf{z})$ at z_d , namely $g(z_d)$, is

$$g(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_w (\bar{\mathbf{Y}} - \bar{\mathbf{X}}\mathbf{B}),$$

where $\bar{\mathbf{X}}$ is a $D \times p$ matrix of the form $(\bar{\mathbf{X}}_1, \dots, \bar{\mathbf{X}}_p)$.

Realistically, we cannot access the whole population. Instead, we can only observe a sample drawn from the population using a certain probability sampling design. Let \mathbf{s} be the set of n sample units with sample $(y_k, \mathbf{x}_k, z_k, w_k)$ for $k \in \mathbf{s}$, where w_k is the sampling weight for unit k . Additionally, we assume that there is complete response so that the inclusion probability is equal to the reciprocal of the sampling weight. We assume further that the bins induced by the distinct values of \mathbf{z} are preserved from the population to the sample. This is appropriate in a variable such as age recorded to age last birthday.

Using the local polynomial regression technique for complex survey data in Bellhouse and Stafford (2001), we use the sampling versions of the objective functions in (4) and (5) as follows,

$$\begin{aligned} &\sum_{d=1}^D \frac{\hat{P}_d}{h} \{ \bar{y}_d - \alpha_0 - \alpha_1(z'_d - z_d), \dots, -\alpha_q(z'_d - z_d)^q \}^2 \\ &\quad \times K\left(\frac{z'_d - z_d}{h}\right) \end{aligned} \quad (9)$$

and

$$\sum_{d'=1}^D \frac{\hat{p}_d}{h} \{ \bar{x}_{jd} - \gamma_0 - \gamma_1(z'_d - z_d), \dots, -\gamma_q(z'_d - z_d)^q \}^2 \times K\left(\frac{z'_d - z_d}{h}\right), \quad (10)$$

where \bar{y} and \bar{x}_j are sample estimators of \bar{Y} and \bar{X}_j and are of the forms $(\bar{y}_1, \dots, \bar{y}_D)^T$ and $(\bar{x}_{j1}, \dots, \bar{x}_{jD})^T$, respectively, and \hat{p}_d is the weighted sample proportion of observations in bin d . Consequently, we have the survey estimator of $m_y(z)$ and $m_j(z)$ at z_d , given by

$$\hat{m}_j(z_d) = \mathbf{e}^T (\mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} \mathbf{Z})^{-1} \mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} \bar{\mathbf{x}}_j \quad (11)$$

and

$$\hat{m}_y(z_d) = \mathbf{e}^T (\mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} \mathbf{Z})^{-1} \mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} \bar{\mathbf{y}},$$

where \mathbf{Z} has the same form as in (6) and $\hat{\mathbf{K}}_{\mathbf{W}}$ is defined as

$$\hat{\mathbf{K}}_{\mathbf{W}} = \frac{1}{h} \text{diag}(\hat{p}_1 K((z_1 - z_d)/h), \dots, \hat{p}_D K((z_D - z_d)/h)).$$

We can also construct the $n \times p$ matrix $\hat{\mathbf{M}}_{\mathbf{x}}$ and $n \times 1$ vector $\hat{\mathbf{M}}_{\mathbf{y}}$ using the same method that we used to construct $\mathbf{M}_{\mathbf{x}}$ and $\mathbf{M}_{\mathbf{y}}$ in equations (7). That is, we use sampling estimators $\hat{m}_j(z_d)$ and $\hat{m}_y(z_d)$ that are shown in (11) to obtain

$$\hat{\mathbf{M}}_{\mathbf{x}} = \begin{pmatrix} \begin{pmatrix} \hat{m}_{x_1}(z_1) & \hat{m}_{x_2}(z_1) & \cdots & \hat{m}_{x_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{m}_{x_1}(z_1) & \hat{m}_{x_2}(z_1) & \cdots & \hat{m}_{x_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \\ \begin{pmatrix} \hat{m}_{x_1}(z_D) & \hat{m}_{x_2}(z_D) & \cdots & \hat{m}_{x_p}(z_D) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{m}_{x_1}(z_D) & \hat{m}_{x_2}(z_D) & \cdots & \hat{m}_{x_p}(z_D) \end{pmatrix} \end{pmatrix}$$

and

$$\hat{\mathbf{M}}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \hat{m}_y(z_1) \\ \vdots \\ \hat{m}_y(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \hat{m}_y(z_D) \\ \vdots \\ \hat{m}_y(z_D) \end{pmatrix} \end{pmatrix}.$$

Let n_d be the number of observations in the d^{th} bin such that $\sum n_d = n$. Similar to $\mathbf{M}_{\mathbf{x}}$ and $\mathbf{M}_{\mathbf{y}}$ in (7), the d^{th} blocks of $\hat{\mathbf{M}}_{\mathbf{x}}$ and $\hat{\mathbf{M}}_{\mathbf{y}}$ are of the dimensions of $n_d \times p$ and $n_d \times 1$, respectively.

Analogous to the population estimating equation (8), the sampling estimating equation for \mathbf{B} is

$$\hat{\mathbf{u}}(\hat{\theta}) = \sum_{k \in s} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k})^T (y_k - \hat{M}_{\mathbf{x}k}) w_k - \sum_{k \in s} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k})^T (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k}) \hat{\mathbf{B}} w_k = \mathbf{0}, \quad (12)$$

where $\hat{\theta}^T = (\hat{\mathbf{B}}^T, \hat{\mathbf{m}}_{\mathbf{x}}(z), \hat{\mathbf{m}}_{\mathbf{y}}(z)^T)$ is the sampling estimator of $\theta^T = (\mathbf{B}^T, \mathbf{m}_{\mathbf{x}}(z), \mathbf{m}_{\mathbf{y}}(z)^T)$. Note that a similar approach was considered by Fuller (1975) and Binder (1983). Nevertheless, the solution to (12) provides the closed form of $\hat{\mathbf{B}}$ as

$$\hat{\mathbf{B}} = ((\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{x}})^T \mathbf{W}_n (\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{x}}))^{-1} (\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{x}})^T \mathbf{W}_n (\mathbf{y} - \hat{\mathbf{M}}_{\mathbf{y}}),$$

where \mathbf{W}_n is an $n \times n$ weight matrix with design weights w_k on the diagonal entry for $k \in s$, \mathbf{y} is an $n \times 1$ vector containing the sample observations of the response variable and \mathbf{x} is an $n \times p$ matrix consisting of the sample observations of the covariates.

Using the sample estimates of \mathbf{B} and denoting $\bar{\mathbf{x}}$ as a $D \times p$ matrix of the form $(\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_p)$, we can obtain the sampling estimate of $g(z_d)$ as

$$\hat{g}(z_d) = \mathbf{e}^T (\mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} \mathbf{Z})^{-1} \mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} (\bar{\mathbf{y}} - \bar{\mathbf{x}} \hat{\mathbf{B}}).$$

Again, if q and h are the same as for $\hat{m}_j(z_d)$, the expression for $\hat{g}(z_d)$ simplifies.

When applying local polynomial regression techniques to obtain the estimators of conditional expectations as well as the arbitrary function $G(\cdot)$, we need to choose an appropriate bandwidth h . Because binning is involved in all aspects of the estimation process and since we assume that bins induced by the distinct values of \mathbf{z} are preserved from the population to the sample, we argue that the same bandwidth should be used for obtaining both the census estimators and the sample estimators. Since we do not have all the observations of the finite population, we use the sample to choose the appropriate band width. In this paper, we adopt the method in Fan and Gijbels (1995), where the authors developed a data-driven bandwidth selector that combines the ideas of the plug-in and the cross-validation methods for the identically and independently distributed data. When applying this data-driven method to our case, criteria, such as the residual sum of squares and mean square error, of the resulting estimates of the conditional expectations are needed. By noting that those criteria depend on the estimated conditional expectations or regression functions and the derivatives of the regression functions, we

can use the objective functions defined in (9) and (10) to obtain not only the survey estimates of regression functions, but also the derivatives of the regression functions. For more details, see Wang (2004).

3. Design properties of sampling estimators

3.1 Notation and assumptions

In showing design properties of the estimators, we follow Särndal, Swensson and Wretman (1992) and Isaki and Fuller (1982) in considering a nested sequence of populations U_v , for $v = 1, 2, \dots$, such that $U_1 \subset U_2 \subset U_3 \subset \dots$. All population quantities, sample sizes and values, and survey estimators are indexed by v . However, for ease of notation we drop v as a subscript for these quantities. We denote the expectation and variance with respect to sampling design as E_p and Var_p , respectively, and in accordance with the above nested populations, we define design-based consistency and asymptotic unbiasedness as in Thompson (1997, page 167).

In what follows, the development of the asymptotic results for the estimators will depend on the asymptotic normality and consistency of the estimates of means and totals. We will not restrict ourselves to specific sampling designs; instead, we assume that all the survey totals that appear in the estimators are of the Horvitz-Thompson type. Hence, the consistency and asymptotic normality of estimators are subject to the standard regularity conditions on the sampling designs for the consistency and normality of Horvitz-Thompson type estimators, which have been studied by Madow (1948), Hájek (1960), Bickel and Freedman (1983), Krewski and Rao (1981) and Shao (1996). The aforementioned literature shares some restrictions on the sampling design. An implication of these restrictions is that no survey weight is disproportionately large, the total number of first stage sampled clusters or primary sampling units is increasing, but with a growing gap between sample and population. In addition, a Liapunov - type condition ensures that the variables z , x and y develop in a regular manner as v tends to infinity.

We will use the result that any vector of estimators of totals from binned data is asymptotically multivariate normal, provided that the conditions in the previous paragraph are met and the number of domains is fixed. This is obtained through application of results in Shao (1996, page 211) and Serfling (1980, page 18). Shao (1996) shows that in this framework any smooth function of estimates of totals is asymptotically normal. An estimate of a domain mean is one such smooth function. Likewise, any linear combination of different domain mean estimates is a smooth function of survey estimates of totals. For our purposes, the

bins form the domains and hence any vector of bin means is asymptotically multivariate normal. The asymptotic result used here depends on having a fixed number of bins. However, it can be incorporated in principle into a theory of the superpopulation parameters, as for example in the approach of Buskirk and Lohr (2005).

Define $\hat{\mathbf{m}}_\xi(\mathbf{z}) = (\hat{\mathbf{m}}_x(\mathbf{z}), \hat{\mathbf{m}}_y(\mathbf{z})^T)^T$ as the survey estimator of $\mathbf{m}_\xi(\mathbf{z}) = (\mathbf{m}_x(\mathbf{z}), \mathbf{m}_y(\mathbf{z})^T)^T$. Using a Taylor linearization technique on (12) and letting ε denote a quantity approaching 0 and as $\hat{\theta}$ approached to θ , we have

$$-\hat{\mathbf{u}}_B(\theta)(\hat{\mathbf{B}} - \mathbf{B}) \doteq \hat{\mathbf{u}}(\theta) + \hat{\mathbf{U}}_\xi(\theta)(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) + \|\hat{\theta} - \theta\| \varepsilon, \quad (13)$$

where $\hat{\mathbf{u}}(\theta)$ is a linear sampling estimator of $\mathbf{u}(\theta)$ in (8) and is of the form

$$\hat{\mathbf{u}}(\theta) = \sum_{k \in s} (\mathbf{x}_k - \mathbf{M}_{xk})^T (y_k - M_{yk}) w_k - \sum_{k \in s} (\mathbf{x}_k - \mathbf{M}_{xk})^T (\mathbf{x}_k - \mathbf{M}_{xk}) \mathbf{B} w_k; \quad (14)$$

$\hat{\mathbf{u}}_B(\theta)$ is the gradient of $\hat{\mathbf{B}}$ obtained from $\hat{\mathbf{u}}(\theta)$; and $\hat{\mathbf{U}}_\xi(\theta)$ is a $p \times (p+1)D$ matrix whose components are the first derivatives of $\hat{\mathbf{u}}(\theta)$ with respect to $\mathbf{m}_\xi(\mathbf{z})$. Denote by $\mathbf{u}_B(\theta)$ and $\mathbf{U}_\xi(\theta)$ the population parameters corresponding to $\hat{\mathbf{u}}_B(\theta)$ and $\hat{\mathbf{U}}_\xi(\theta)$, respectively.

In addition to the aforementioned regularity conditions, we impose the following conditions, letting \mathcal{N} denote a neighbourhood of the true value of the parameters of interest.

- C1. $\lim_{v \rightarrow \infty} \mathbf{u}(\theta)/N$ exists and is finite for all θ and \mathcal{N} .
- C2. $\lim_{v \rightarrow \infty} \mathbf{u}_B(\theta)/N = \mathbf{H}_B$ and \mathbf{H}_B is of full rank and is invertible for all θ and \mathcal{N} .
- C3. $\lim_{v \rightarrow \infty} \mathbf{U}_\xi(\theta)/N = \mathbf{H}_\xi(\theta)$ and $\mathbf{H}_\xi(\theta)$ has a finite determinant for all θ and \mathcal{N} .
- C4. $\lim_{v \rightarrow \infty} n \text{Var}_p(\hat{\mathbf{u}}(\theta)/N) = \mathbf{V}(\hat{\mathbf{u}}(\theta))$ where Var_p is the design-based variance and $\mathbf{V}(\hat{\mathbf{u}}(\theta))$ is a positive-definite variance matrix for all θ and \mathcal{N} .
- C5. $\lim_{v \rightarrow \infty} N_d/N = \omega_d$ and $\lim_{v \rightarrow \infty} n/N = f$ with both ω_d and f are constants between 0 and 1.
- C6. Let $\mathbf{A}_d = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}_W \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_W$ be the population smoothing matrix; then $\lim_{v \rightarrow \infty} \mathbf{A}_d$ exists and is finite for $d = 1, \dots, D$.
- C7. $\lim_{v \rightarrow \infty} n \text{Var}_p(\hat{\mathbf{m}}_\xi(\mathbf{z})) = \mathbf{V}(\hat{\mathbf{m}}_\xi(\mathbf{z}))$.
- C8. Matrices of population values $\mathbf{Z}^T \mathbf{K}_W \mathbf{Z}$ and $\mathbf{u}_B(\theta)$ are invertible, as well as their sampling estimators $\mathbf{Z}^T \hat{\mathbf{K}}_W \mathbf{Z}$ and $\hat{\mathbf{u}}_B(\hat{\theta})$.

3.2 Asymptotic properties of $\hat{\mathbf{B}}$

The proofs of all lemmas and theorems in this and the following section may be found in the Appendix. From the Taylor linearization results in (13), we know that the properties of $\hat{\mathbf{B}}$ are dependent on those of $\hat{\mathbf{u}}(\theta)$, $\hat{\mathbf{u}}_{\mathbf{B}}(\theta)$, $\hat{\mathbf{U}}_{\xi}(\theta)$ and $\hat{\mathbf{m}}_{\xi}(\mathbf{z})$; their properties are stated in the following two Lemmas.

Lemma 1. *If conditions C1 – C4 are satisfied, we have as $v \rightarrow \infty$:*

- 1) $\sqrt{n}(\hat{\mathbf{u}}(\theta) - \mathbf{u}(\theta))/N \rightarrow N(0, V(\hat{\mathbf{u}}(\theta)))$;
- 2) $|\hat{\mathbf{u}}_{\mathbf{B}}(\theta) - \mathbf{u}_{\mathbf{B}}(\theta)|/N$ and $|\hat{\mathbf{U}}_{\xi}(\theta) - \mathbf{U}_{\xi}(\theta)|$ converge to 0 in probability for θ and \mathcal{N} ;
- 3) $|\hat{\mathbf{u}}(\theta) - \mathbf{u}(\theta)|/N$ converges to zero in probability.

Lemma 2. *Under conditions C5 to C7, $\sqrt{n}(\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})) = O_p(1)$.*

Building on Lemmas 1 and 2, we have the asymptotic normality of $\hat{\mathbf{B}}$ in Theorem 1.

Theorem 1. *Under conditions C1 to C7, assuming the parameter space contains a neighbourhood of the parameter of interest, we have as v goes to infinity:*

- 1) $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) \rightarrow N(0, V(\hat{\mathbf{B}}))$ where $V(\hat{\mathbf{B}}) = \lim_{v \rightarrow \infty} n \text{Var}_p(\hat{\mathbf{B}})$;
- 2) $|\hat{\mathbf{B}} - \mathbf{B}|$ converges to zero in probability.

To obtain approximate moments for $\hat{\mathbf{B}}$, we take expectations on both sides of equation (13), which yields

$$\begin{aligned} E_p(-\hat{\mathbf{u}}_{\mathbf{B}}(\theta)(\hat{\mathbf{B}} - \mathbf{B})) &\doteq E_p(\hat{\mathbf{u}}(\theta)) \\ &+ E_p\{\hat{\mathbf{U}}_{\xi}(\theta)[\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})]\} \\ &+ E_p(\|\hat{\theta} - \theta\|)\varepsilon. \end{aligned} \quad (15)$$

The assumption that the second moments of the estimates are bounded makes the last term of equation (15) vanish in the limit. Following along the lines of Binder (1983), we have

$$\begin{aligned} E_p(-\hat{\mathbf{u}}_{\mathbf{B}}(\theta))E_p((\hat{\mathbf{B}} - \mathbf{B})) &\doteq E_p(\hat{\mathbf{u}}(\theta)) \\ &+ E_p(\hat{\mathbf{U}}_{\xi}(\theta))E_p\{\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})\}. \end{aligned}$$

The survey totals that define the vector $\hat{\mathbf{u}}(\theta)$ and matrix $\hat{\mathbf{u}}_{\mathbf{B}}(\theta)$ are Horvitz-Thompson-type estimators and they are unbiased (Thompson 1997). Hence, $E_p(\hat{\mathbf{u}}(\theta)) = \mathbf{u}(\theta)$ and $E_p(\hat{\mathbf{u}}_{\mathbf{B}}(\theta)) = \mathbf{u}_{\mathbf{B}}(\theta)$. Since $\mathbf{u}(\theta)$ is the estimating equation for the partial linear coefficients defined in (8), it is equal to a $1 \times p$ zero vector. Further, it has been shown in Bellhouse and Stafford (2001) that $\hat{\mathbf{m}}_{\xi}(\mathbf{z})$ is an asymptotically unbiased estimator of $\mathbf{m}_{\xi}(\mathbf{z})$. Hence, $-\mathbf{u}_{\mathbf{B}}(\theta)E_p((\hat{\mathbf{B}} - \mathbf{B})) \doteq 0$, or, based on the conditions that $\mathbf{u}_{\mathbf{B}}(\theta)$ is invertible and $\mathbf{u}_{\mathbf{B}}(\theta)^{-1}$ is finite, we have $E_p(\hat{\mathbf{B}}) \doteq \mathbf{B}$.

Taking the variance of both sides of equation (13) and using the approximated variance-covariance matrices of

$\hat{\mathbf{u}}(\theta)$ and $\hat{\mathbf{m}}_{\xi}(\mathbf{z})$, we obtain the asymptotic variance of $\hat{\mathbf{B}}$ as

$$\begin{aligned} \text{Var}_p(\hat{\mathbf{B}}) &\doteq \mathbf{u}_{\mathbf{B}}(\theta)^{-1} \\ &(\text{Var}_p(\hat{\mathbf{u}}(\theta)) + \mathbf{U}_{\xi}(\theta)(\mathbf{A}(\mathbf{J} \otimes \text{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}))\mathbf{A}^T)\mathbf{U}_{\xi}(\theta)^T \\ &+ 2(I_p \otimes \ell)(\ell \otimes \mathbf{C})\mathbf{A}^T\mathbf{U}_{\xi}(\theta)^T)(\mathbf{u}_{\mathbf{B}}(\theta)^T)^{-1}, \end{aligned} \quad (16)$$

where $\text{Var}_p(\hat{\mathbf{u}}(\theta))$ is a $p \times p$ matrix composed of variances of totals in the vector $\hat{\mathbf{u}}(\theta)$ and $\text{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is the variance-covariance matrix of the binned means of the parametric covariates and the response variable. The matrices \mathbf{J} and ℓ are the $D \times D$ unit matrix and the $1 \times D$ unit vector, respectively. Finally, we have

$$\mathbf{A} = \begin{pmatrix} I_{p+1} \otimes \mathbf{A}_1 & 0 & 0 & 0 \\ 0 & I_{p+1} \otimes \mathbf{A}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & I_{p+1} \otimes \mathbf{A}_D \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} \text{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{x}}_1) & \cdots & \text{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{x}}_p) & \text{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{y}}) \\ \vdots & & \vdots & \vdots \\ \text{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{x}}_1) & \cdots & \text{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{x}}_p) & \text{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{y}}) \end{pmatrix},$$

where, for $j = 1, \dots, p$, $\hat{\mathbf{t}}_j$ is a $D \times 1$ vector whose d^{th} entry is $\sum_{k \in s_d} w_{jk} u_{jk}(\theta)$ and $\mathbf{A}_d = \mathbf{e}^T(\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_w$ for $d = 1, \dots, D$.

Replacing θ , $\text{Var}_p(\hat{\mathbf{u}}(\theta))$, $\text{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, \mathbf{A} and \mathbf{C} by their sample estimators, we have the survey estimator of the variance of $\hat{\mathbf{B}}$:

$$\begin{aligned} \widehat{\text{Var}}_p(\hat{\mathbf{B}}) &= \hat{\mathbf{u}}_{\mathbf{B}}^{-1}(\hat{\theta}) \\ &(\widehat{\text{Var}}_p(\hat{\mathbf{u}}(\hat{\theta})) + \hat{\mathbf{U}}_{\xi}(\hat{\theta})(\hat{\mathbf{A}}(\mathbf{J} \otimes \widehat{\text{Cov}}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}))\hat{\mathbf{A}}^T)\hat{\mathbf{U}}_{\xi}(\hat{\theta})^T \\ &+ 2(I_p \otimes \ell)(\ell \otimes \hat{\mathbf{C}})\hat{\mathbf{A}}^T\hat{\mathbf{U}}_{\xi}(\hat{\theta})^T)(\hat{\mathbf{u}}_{\mathbf{B}}^T(\hat{\theta}))^{-1}, \end{aligned}$$

where $\hat{\mathbf{A}}$ is the survey estimator of \mathbf{A} and is composed of $\hat{\mathbf{A}}_d = \mathbf{e}^T(\mathbf{Z}^T \hat{\mathbf{K}}_w \mathbf{Z})^{-1} \mathbf{Z}^T \hat{\mathbf{K}}_w$.

3.3 Asymptotic properties of $\hat{g}(\cdot)$

Define $\bar{\mathbf{r}} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}$ to be the sample estimator of $\bar{\mathbf{R}} = \bar{\mathbf{Y}} - \bar{\mathbf{X}}\mathbf{B}$. A linearization around the population parameters, as well as design unbiasedness of domain means and $\hat{\mathbf{B}}$, leads to the asymptotic design unbiasedness of $\bar{\mathbf{r}}$. Reexpressing $\hat{g}(z_d)$, we have $\hat{g}(z_d) = \hat{\mathbf{A}}_d \bar{\mathbf{r}}$. In $\hat{\mathbf{A}}_d$, we can expand $(\mathbf{Z}^T \hat{\mathbf{K}}_w \mathbf{Z})^{-1}$ using the Taylor series expansion that $(\mathbf{I} + \mathbf{G})^{-1} = \mathbf{I} - \mathbf{G} + \mathbf{G}^2 - \dots$ given that \mathbf{G} is a symmetric and invertible matrix. Using the first two terms of the expansion, we can show that $E_p(\hat{\mathbf{A}}_d)$ is approximately \mathbf{A}_d . Hence, we have the asymptotic design

unbiasedness of $\hat{g}(z_d)$. With the same technique, the approximate asymptotic design-based variance of $\hat{g}(z_d)$ is obtained as

$$\text{Var}_p(\hat{g}(z_d)) = \mathbf{A}_d \text{Var}_p(\bar{\mathbf{r}}) \mathbf{A}_d^T,$$

where, given that $\mathbf{Q} = (1, -B_1, \dots, -B_p)$,

$$\begin{aligned} \text{Var}_p(\bar{\mathbf{r}}) &\doteq (\mathbf{Q} \otimes \mathbf{I}_D) \text{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}) (\mathbf{Q} \otimes \mathbf{I}_D)^T \\ &\quad + \bar{\mathbf{X}} \text{Var}_p(\hat{\mathbf{B}}) \bar{\mathbf{X}}^T \\ &\quad - 2(\mathbf{Q} \otimes \mathbf{I}_D) \text{Cov}_p(\bar{\mathbf{y}}, \hat{\mathbf{B}}) \bar{\mathbf{X}}^T \\ &\quad - \sum_{j=1}^p 2(\mathbf{Q} \otimes \mathbf{I}_D) \text{Cov}_p(\bar{\mathbf{x}}_j, \hat{\mathbf{B}}) \bar{\mathbf{X}}^T. \end{aligned}$$

Given the estimated variance of $\bar{\mathbf{r}}$, namely

$$\begin{aligned} \widehat{\text{Var}}_p(\bar{\mathbf{r}}) &= (\hat{\mathbf{Q}} \otimes \mathbf{I}_D) \widehat{\text{Cov}}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}) (\hat{\mathbf{Q}} \otimes \mathbf{I}_D)^T \\ &\quad + \bar{\mathbf{x}} \widehat{\text{Var}}_p(\hat{\mathbf{B}}) \bar{\mathbf{x}}^T \\ &\quad - 2(\mathbf{Q} \otimes \mathbf{I}_D) \widehat{\text{Cov}}_p(\bar{\mathbf{y}}, \hat{\mathbf{B}}) \bar{\mathbf{x}}^T \\ &\quad - \sum_{j=1}^p 2(\hat{\mathbf{Q}} \otimes \mathbf{I}_D) \widehat{\text{Cov}}_p(\bar{\mathbf{x}}_j, \hat{\mathbf{B}}) \bar{\mathbf{x}}^T, \end{aligned}$$

the estimated variance of $\hat{g}(z_d)$ is $\widehat{\text{Var}}_p(\hat{g}(z_d)) = \hat{\mathbf{A}}_d \widehat{\text{Var}}_p(\bar{\mathbf{r}}) \hat{\mathbf{A}}_d^T$.

The asymptotic normality of $\hat{g}(\cdot)$ is also dependent on the normality of $\bar{\mathbf{r}}$, which is shown in the following Lemma.

Lemma 3. Under conditions C1 to C7 and assuming that the dimension of $\bar{\mathbf{r}}$ is finite, we have as v goes to infinity

$$\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}(\bar{\mathbf{r}})),$$

where

$$\mathbf{V}(\bar{\mathbf{r}}) = \lim_{v \rightarrow \infty} n \text{Var}_p(\bar{\mathbf{r}}).$$

Based on the asymptotic normality developed in Lemma 3 and the estimator of the variance of $\hat{g}(z_d)$, we establish the asymptotic properties of $\hat{g}(z_d)$ in the following Theorem.

Theorem 2. Under conditions C1 to C7, we have as v goes to infinity:

- 1) $|\hat{g}(z_d) - g(z_d)| \rightarrow 0;$
- 2) $(\hat{g}(z_d) - g(z_d)) / \sqrt{\widehat{\text{Var}}_p(\hat{g}(z_d))} \xrightarrow{d} N(0, 1).$

4. Simulation studies

4.1 Design of experiment

The simulation study implemented here was designed to illustrate the theoretical results in Theorems 1 and 2. We generated the data in a two-step process that mimicked a superpopulation approach to sampling. First, we generated the finite population and then the sample was selected from it. In particular, we considered a finite population of $L = 500$ clusters with $M (= M_i) = 20,000$ in each. The population observations for the measurement of interest y_{ij} were obtained from the model

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \\ &\quad + 0.5 \exp\left(\frac{z_{ij} - 40}{10}\right) + \mu_i + \varepsilon_{ij} \end{aligned} \quad (17)$$

for $i = 1, \dots, L$ and $j = 1, \dots, M$ where the error terms μ_i and ε_{ij} are mutually independent with $\mu_i \sim N(0, \sigma_\mu^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. We set $\sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$ so that the intraclass correlation coefficient is $\rho = \sigma_\mu^2 / \sigma^2$. Among the covariates in the model, both x_{1ij} and x_{2ij} were treated as the parametric linear part of the model and z_{ij} as the nonparametric part. We generated the x_{1ij} from the Bernoulli(1/2) distribution and the x_{2ij} from the Uniform(0, 1) distribution. The z_{ij} were generated from the age distribution of the Canadian population (according to the 1996 census) for the 18 to 64 age range and were independent of the error terms. Results for the values $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$, $\sigma^2 = 3$ and $\rho = 0, 0.2, 0.5$ are reported in this study. A two-stage sampling design, with $l = (10, 25, 50, 100)$ clusters chosen at random from L and $m (= 1,000)$ secondary sampling units chosen at random from each cluster of size M , was used for the study. For each sample size and value of ρ , the simulation was repeated 300 times. At the population level, we applied the bandwidth selection method from Fan and Gijbels (1995) and determined that the bandwidths for estimating the conditional expectations of X_1 and X_2 on z were 1.2 and 1.5 respectively. When smoothing the residuals to estimate $g(z)$, the bandwidth was 0.6.

4.2 Results

Using the generated finite population, we found that the census estimates were $B_1 = 2.01$ and $B_2 = 3.00$. To check the design unbiasedness and efficiency of $\hat{\mathbf{B}}$, we calculated the simulated squared bias (Bias^2), which is the square of the difference between the average of the simulated estimates and the census estimates. In addition, the ratio of the average variance estimates to the simulated variance of each estimator of a linear coefficients (RVar) is presented to show the validity of the variance estimator $\text{Var}_p(\hat{\mathbf{B}})$. To

evaluate the normality of $\hat{\mathbf{B}}$, we standardized the estimates of linear coefficients using the empirical standard deviation and population value of \mathbf{B} and graphed the quantile - quantile plots of the standardized values.

Applying the semiparametric technique in Speckman (1988) to the model (17), we obtained census estimates $g(z)$ for $z = 18, \dots, 64$. To evaluate the design accuracy of $\hat{g}(z)$, we took the difference between $\hat{g}(z)$ and $g(z)$ at each distinct point. The average of the squares of the differences over 47 distinct values of z is then reported as $ABias^2$. Two mean square errors were computed to check the design efficiency of $\hat{g}(z)$ and convergence of $\widehat{Var}_p(\hat{g}(z))$. One of the mean square errors is the average of the estimates of the integrated mean square error (AIMSE), which is obtained by first summing the $\widehat{Var}_p(\hat{g}(z))$ over $z = 18, \dots, 64$ for each simulation and then taking the average of the sums over the total number of simulations. The simulated integrated mean square error (IMSE) is another mean square error and was computed by summing up the simulated mean square error at each distinct point of z . The average of the ratios of the simulated mean of $\widehat{Var}_p(\hat{g}(z))$ to the simulated variance of $\hat{g}(z)$ (Reff) shows the convergence of $\widehat{Var}_p(\hat{g}(z))$. In addition, we computed the coverage of the pointwise 95% confidence interval at each distinct point of z .

The results on the properties of $\hat{\mathbf{B}}$, $\widehat{Var}_p(\hat{\mathbf{B}})$, $\hat{g}(z)$ and $\widehat{Var}_p(\hat{g}(z))$ are found in Tables 1 and 2 and Figures 2 and 3. Tables 1 and 2 show information about accuracy and precision of the simulated estimates of $\hat{\mathbf{B}}$ and $\hat{g}(\cdot)$. Figure 2 gives the quantile-quantile plots of the sample standardized value of \hat{B}_2 . Note that the quantile-quantile plots for \hat{B}_1 behave in a similar way to those for \hat{B}_2 . Figure 3 graphs the coverage of the 95% confidence intervals for $g(\cdot)$. In Figures 2 and 3, we only report the cases where $l = 10, 25, 100$ and $\rho = 0, 0.5$. The overall performance of the estimators agrees with the theory in Theorems 1 and 2.

Table 1 confirms the design unbiasedness of $\hat{\mathbf{B}}$. It also shows that as the sample size increases, the performances of

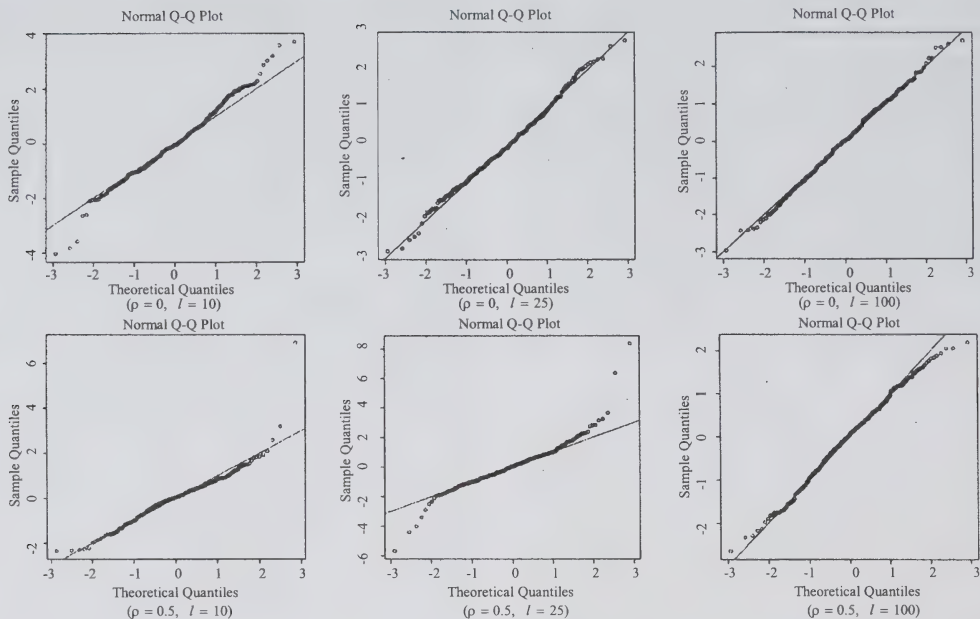
the estimates of the linear coefficients improve for all the error structures. In particular, the squared bias and variance of $\hat{\mathbf{B}}$ decreases as the number of primary samples increases. The estimated variance of $\hat{\mathbf{B}}$ gets closer to the simulated variance of $\hat{\mathbf{B}}$ as the sample size increases; this confirms the consistency of the variance estimates of $\hat{\mathbf{B}}$. Comparing the variances and biases of $\hat{\mathbf{B}}$ in the cases that $\rho = 0.2$ and $\rho = 0.5$ to the case where $\rho = 0$, we found that the intraclass correlation (cluster effects) did not affect the performance of $\hat{\mathbf{B}}$. This may be because the within cluster sample size was large.

Observing Figure 2, we find that both the number of primaries sampled and the cluster effect play some role in the normality of $\hat{\mathbf{B}}$. In particular, when the primary sample size is low, for instance $l = 10$, normality of the standardized $\hat{\mathbf{B}}$ shows some deviation from the theory for both $\rho = 0$ and $\rho = 0.5$. When l increases to 25, we find that performance of $\hat{\mathbf{B}}$ for $\rho = 0$ starts to recover whereas, for $\rho = 0.5$, there is no improvement until $l = 100$. Empirically, this finding suggests that when the number of clusters is low, we should not rely on the theoretical normality of the estimates of the coefficient; instead, we may want to use t distribution to carry out the inference.

As for the results of the nonparametric part of the estimation, Table 2 shows that the average estimated integrated mean square errors are very close to the simulated integrated mean square errors for all the sample sizes and error structures. Design unbiasedness is again confirmed with the average squared bias ($ABias^2$). The values of average ratio of the estimated variance to the simulated variance (RVar), which are close to 1 for all cases, are in line with the design consistency of the estimator of the variance of $\hat{g}(z)$. The integrated mean square errors of $\hat{g}(\cdot)$ are influenced by the intraclass correlations. This can be shown by the fact that the approach to zero of both integrated mean square error and average estimated integrated mean square error is slower in the cases where $\rho = 0.2$ and $\rho = 0.5$ than in the case where $\rho = 0$.

Table 1
Simulation results for point estimators of $\hat{\mathbf{B}}$

	<i>l</i>	$\rho = 0$			$\rho = 0.2$			$\rho = 0.5$		
		Bias ² ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Bias ² ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Bias ² ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar
\hat{B}_1	10	5.77	1.07	1.13	3.12	1.1	1.01	0.23	1.19	1.33
	25	9.97	0.46	1.07	0.38	0.44	1.08	0.30	0.53	0.98
	50	0.54	0.21	1.08	0.13	0.27	0.93	0.026	0.21	1.18
	100	0.22	0.13	0.96	0.019	0.11	1.06	0.039	0.13	0.98
\hat{B}_2	10	0.36	3.32	1.13	1.54	3.74	0.92	1.26	3.5	1.78
	25	0.64	1.31	1.10	2.40	1.34	1.06	0.14	1.42	1.03
	50	0.31	0.75	0.94	1.27	0.85	0.94	0.16	0.76	0.97
	100	0.15	0.38	0.94	1.11	0.38	0.98	0.072	0.33	1.03

Figure 2 Quantile – quantile plots for standardized \hat{B}_2 Table 2
Bias and efficiency of $\hat{g}(z)$

ρ	l	AIMSE	IMSE	ABias ² ($\times 10^{-5}$)	RVar
0	10	0.37	0.42	5.29	1.27
	25	0.15	0.17	3.20	1.10
	50	0.074	0.086	3.29	1.09
	100	0.037	0.044	2.34	1.08
0.2	10	2.95	3.25	6.13	0.91
	25	1.22	1.17	3.71	1.04
	50	0.74	0.54	2.34	1.0
	100	0.26	0.27	7.08	0.98
0.5	10	8.143	8.877	3.73	0.92
	25	3.155	3.073	6.56	1.03
	50	1.461	1.599	2.86	1.15
	100	0.659	0.607	3.59	1.09

The coverage of the point-wise 95% confidence intervals for $g(\cdot)$ in Figure 3 varies between 85% and 96%. The coverage improves as the sample size increases. The performance of $\hat{g}(\cdot)$ is, however, more sensitive to the

lower effective sample size caused by the intraclass correlation. In particular, the coverages of the 95% confidence intervals in the cases of $\rho = 0.2$ and $\rho = 0.5$ are smaller than the 95% nominal confidence level when $l = 10$. The coverage improves as the number of primary sampling units increases for the cases of $\rho = 0$ and $\rho = 0.2$. For $\rho = 0.5$, the undercoverage is still present when the sample size increases to 100. It is also seen that at $z = 18$ or 64 , the coverages are higher even than the nominal level; this is because the boundary effect of the local polynomial regression estimation causes larger bias at the two boundaries of the data. For $\rho = 0.5$, the effective sample size is low so that the boundary effect becomes severe, creating the downward spikes at 18 and 63.

It is worth pointing out that although the size of the primary sampling units is large (1,000), the sampling fraction is very small (0.05). Hence, this performance of the estimates would not change even though the size of the primary sampling units is small.

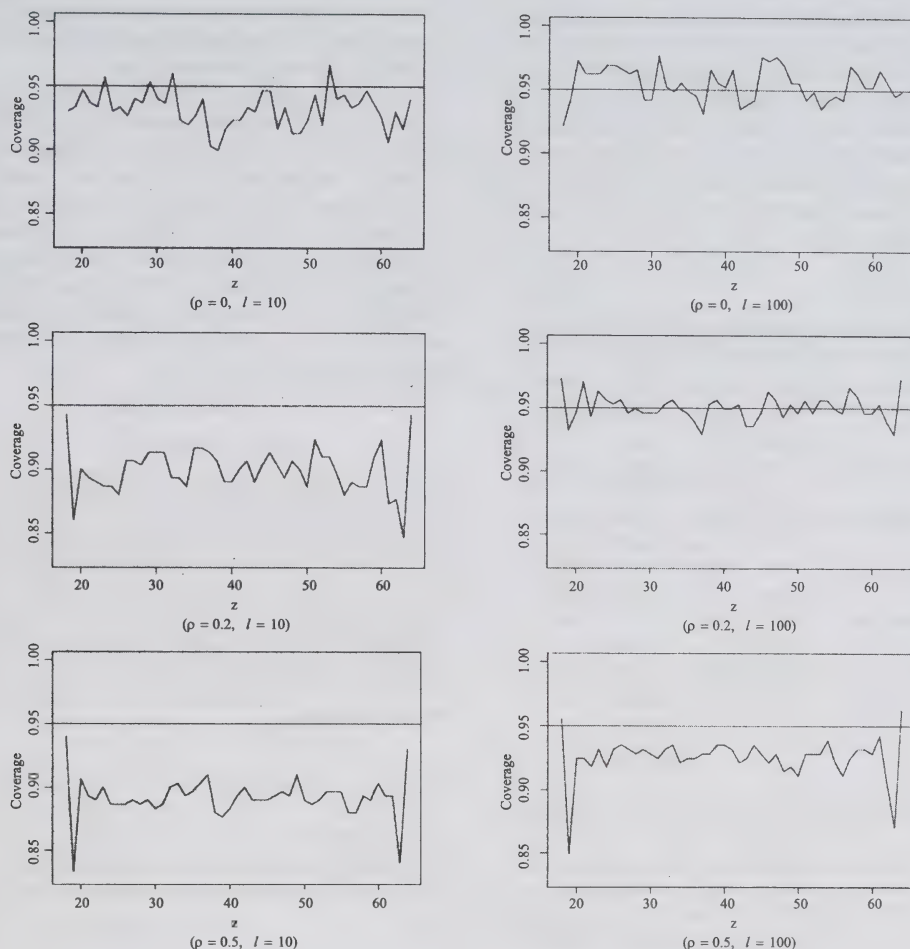


Figure 3 Coverage of the 95% confidence point-wise intervals for $g(z)$

5. Empirical illustrations

We now return to the example introduced in Section 1. For the purpose of illustrating the partial linear model, we examine the effects of age, gender, smoking status and physical activity on the body mass index (BMI) and the desired body mass index (DBMI). Similar to the measure BMI, DBMI is a derived variable for the question asking about the desired weight of a person. Since people stop growing for the age group for which we are interested, we use the actual height when calculating DBMI. We use age as the nonparametric covariate and treat the other factors as

discrete variables. Since there are only 47 distinct points in the age variable, we bin the data set according to age. The bin size is set to unity such that there are 47 bins, with midpoints being 18, 19, ..., 64. Among all the categorical explanatory variables, gender has two levels, male = 1 and female = 0; smoking status includes levels such as former smoker = 0, never smoked = 1, occasional smoker = 2, daily smoker = 3; and physical activeness is divided into three levels: active = 0, moderately active = 1 and inactive = 2. The regression models are (a) $BMI = g_1(\text{age}) + \mathbf{XB}_1 + \varepsilon_1$ and (b) $DBMI = g_2(\text{age}) + \mathbf{XB}_2 + \varepsilon_2$, where \mathbf{X} is the design matrix including all the indicator variables.

Table 3 lists all the survey estimates of the linear coefficients in the models (a) and (b). On comparing BMI by gender, we found that male BMI is higher. Using former smoker as the base category, the coefficients of smoking status are all negative and significant, which suggests that former smokers tend to be heavier than people with other types of smoking status. The estimates also indicate that inactive people have higher BMI. With respect to the DBMI, p -values suggest that most of the life style related factors are not significant.

Table 3
Results for semiparametric regression models (a) and (b) (Values in the parenthesis are the standard errors)

Variable	\hat{B}_1	p -value	\hat{B}_2	p -value
Gender	1.45 (0.05)	0.00	2.80 (0.05)	0.00
Never Smoked	-0.45 (0.10)	0.00	-0.06 (0.06)	0.34
Occasional Smoker	-0.31 (0.17)	0.04	-0.00 (0.10)	0.96
Daily Smoker	-0.61 (0.09)	0.00	-0.12 (0.06)	0.03
Moderately Active	-0.33 (0.09)	0.00	-0.07 (0.06)	0.24
Active	-0.50 (0.09)	0.00	-0.14 (0.09)	0.07

In Figure 4, the estimated functions of age, $\hat{g}_1(\text{Age})$ and $\hat{g}_2(\text{Age})$, and their confidence bands are plotted versus different ages. It is found that, in both cases, the BMI and the DBMI are increasing functions of age.

Figure 5 gives the estimated functions of age, $\hat{g}_1(\text{Age})$ and $\hat{g}_2(\text{Age})$, for active and moderately active people. If we look at the age effect for female and male separately, we find that for females who are either active or moderately active on average the DBMI is lower than the BMI, whereas males with the same intensity of physical activity desire to be heavier before age 21. In addition, we also compare the age trends in the BMI and the DBMI for both the females and males. Due to the inconsistency between the female and male trends, we can conclude that there are interactions between the gender factor and age.

6. Conclusion

With the assistance of a partial linear model, we extend semi-parametric regression techniques to complex survey data. Asymptotic properties of the survey estimators are developed. Computation of the variance estimates of both the linear coefficients and the regression function rely on the variance estimates of survey totals and means. Provided that we obtain the required variance estimates of survey totals and means, we can apply this method using standard statistical packages.

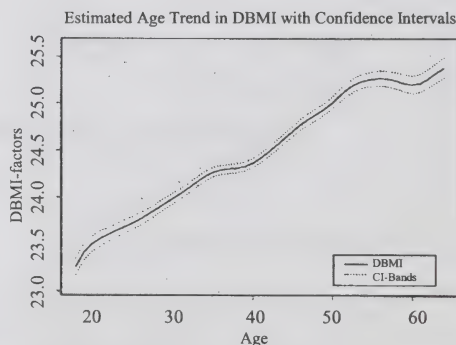
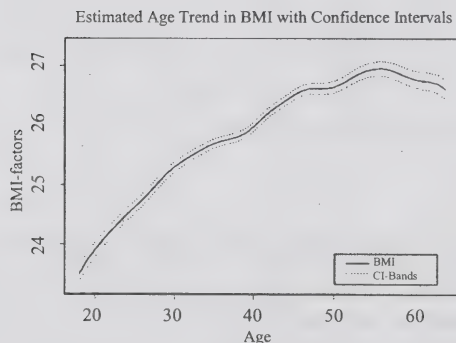


Figure 4 Estimated age trends in BMI and DBMI with 95% pointwise confidence intervals

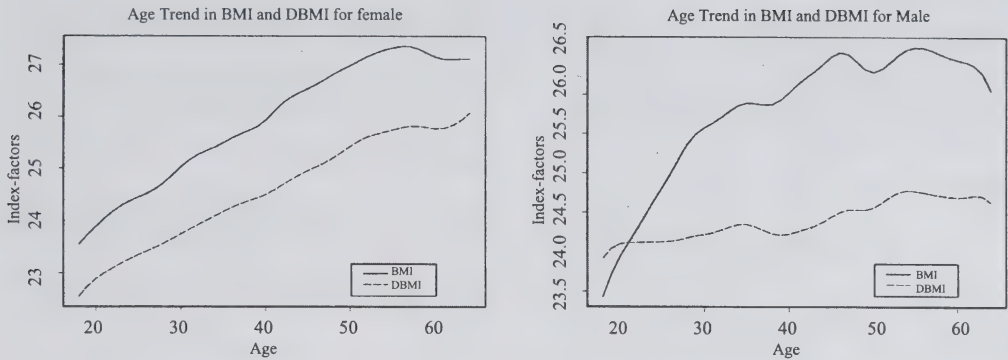


Figure 5 A comparison of estimated age trends in BMI to DBMI for both female and male who are active and moderately active

In the partial linear working model, we assume that there is no interaction between the parametric component and the nonparametric component. However, the empirical example of the age trends of the body mass index has illustrated that this assumption needs to be checked. In future work, we will relax the no interaction assumption. A direct approach to modelling interaction terms is to let the nonparametric component appear linearly in the interaction term. That is, we define the partial linear model as

$$\mathbf{y} = G(\mathbf{z}) + \mathbf{X}\beta + \mathbf{X}H(\mathbf{z}) + \varepsilon.$$

By testing the departure of $H(\mathbf{z})$ from zero, we can detect the existence of interaction.

When estimating conditional expectation on the nonparametric components for indicator discrete random variables, we propose to use generalized linear or additive models to conduct the estimation.

Appendix

A.1 Proof of lemma 1

Observing that entries of $\hat{\mathbf{u}}(\theta)$, $\hat{\mathbf{u}}_B(\theta)$ and $\hat{\mathbf{U}}_\xi(\theta)$ are either sample totals or ratios of sample totals, we can apply Lemmas 1.2.5 and 1.2.6 in Wang (2004) to establish this Lemma.

A.2 Proof of lemma 2

Each entry of $\hat{\mathbf{m}}_\xi(\mathbf{z})$ is just an estimated regression function with the local polynomial technique developed by Bellhouse and Stafford (2001). Theorem 2.2.1 in Wang (2004) shows that $\hat{\mathbf{m}}_\xi(\mathbf{z})$ is root- n consistent. Hence, since the dimension of $\hat{\mathbf{m}}_\xi(\mathbf{z})$ is finite, we can show that $\sqrt{n}(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z}))$ is bounded in probability.

A.3 Proof of theorem 1

Since for the true θ , we have $\mathbf{u}(\theta) = \mathbf{0}$, we can rewrite equation (13) as follows:

$$\begin{aligned} & -\frac{\sqrt{n}\hat{\mathbf{u}}_B(\theta)}{N}(\hat{\mathbf{B}} - \mathbf{B}) \doteq \\ & \left(\frac{\sqrt{n}}{N}(\hat{\mathbf{u}}(\theta) - \mathbf{u}(\theta)) + \hat{\mathbf{U}}_\xi(\theta)\frac{\sqrt{n}}{N}(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) \right) \\ & + \frac{\sqrt{n}}{N}\|\hat{\theta} - \theta\|\varepsilon. \end{aligned}$$

The standard argument in Rao (1973, page 387) yields

$$\sqrt{n}/N\|\hat{\theta} - \theta\|\varepsilon \xrightarrow{P} 0.$$

Using the condition that the sampling fraction $f = n/N$ is constant as n goes to infinity, we have,

$$\begin{aligned} \sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) &= -\left(\frac{\hat{\mathbf{u}}_B(\theta)}{N}\right)^{-1} \\ & \left(\frac{\sqrt{n}}{N}(\hat{\mathbf{u}}(\theta) - \mathbf{u}(\theta)) + \hat{\mathbf{U}}_\xi(\theta)\frac{f\sqrt{n}}{n}(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) \right). \end{aligned}$$

Following from the results in Lemma 1, both $(\hat{\mathbf{u}}_B(\theta)/N)^{-1}$ and $\hat{\mathbf{U}}_{\mathbf{m}_\xi}(\theta)$ converge to their population values in probability. Lemma 2 indicates that the vector $\sqrt{n}(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) = O_p(1)$. Thus, $(f\sqrt{n}/n)(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z}))$ converges to a zero vector in probability as n goes to infinity. Finally, from the normality of $\sqrt{n}(\hat{\mathbf{u}}(\theta) - \mathbf{u}(\theta))/N$ stated in Lemma 1, we use the Slutsky Theorem to show the asymptotic normality of $\hat{\mathbf{B}}$.

A.4 Proof of lemma 3

Given that $\bar{\mathbf{r}} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}$, we have $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) = \sqrt{n}[(\bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}) - \bar{\mathbf{R}}]$. Based on Theorem 1, we know that in the limit as v goes to infinity, $\hat{\mathbf{B}}$ converges to \mathbf{B} in probability. Hence, we have $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) \doteq \sqrt{n}[(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B}) - \bar{\mathbf{R}}]$. The d^{th} entry of $(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B})$ is,

$$\bar{y}_d - \bar{x}_{1d}B_1 - \cdots - \bar{x}_{pd}B_p = \frac{1}{N_d} \sum_{k \in s_d} w_k (y_k - x_{1k}B_1 - \cdots - x_{pk}B_p).$$

That is, $(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B})$ is merely a vector of estimated binned means. Using the result from Shao (1996) on functions of sample means and “Cramer-Wold device” results found in Serfling (1980, page 18), we see that $\sqrt{n}(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B} - \bar{\mathbf{R}})$ converges to a random vector distributed normally. Thus, using this indirect Slutsky idea, we have proved the normality of $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) = \sqrt{n}[(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B}) - \bar{\mathbf{R}}]$.

A.5 Proof of theorem 2

The proof follows the same argument that $\hat{g}(z_d)$ is a function of domain mean and proportions as does in the proof of theorem 2.2.1 in Wang (2004).

Acknowledgements

This work is supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors are grateful to Mary Thompson for her valuable comments and suggestions on the early draft of this paper. The authors also wish to thank the Associate editor and two referees for their very helpful comments.

References

- Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex survey. *Statistica Sinica*, 9, 407-424.
- Bellhouse, D.R., and Stafford, J.E. (2001). Local polynomial regression in complex survey. *Survey Methodology*, 27, 197-203.
- Bickel, P.J., and Freedman, D.A. (1983). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.
- Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.
- Buskirk, D.T., and Lohr, L.S. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.
- Fan, J., and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society, series B*, 57, 371-394.
- Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā, C*, 37, 117-132.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54, 127-138.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudományos Akadémia Budapest Matematikai Kutató Intézet Közleményei*, 5, 361-374.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.
- Jones, M.C. (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84, 733-741.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balance repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.
- Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535-545.
- Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.
- Ontario Health Survey (1992). *Ontario Health Survey: User's Guide*. Ministry of Health, Toronto, Ontario, Canada.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd Ed.). New York: John Wiley & Sons, Inc.
- Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56, 931-954.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Serfling, R.J. (1980). *Approximation Theorem of Mathematical Statistics*. New York: John Wiley & Sons, Inc.
- Shao, J. (1996). Resampling methods in sample survey. *Statistics*, 27, 203-254.
- Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, 50, 413-436.
- Thompson, M.E. (1997). *Theory of Sample Survey* (1st Ed.). New York: Chapman and Hall.
- Wang, Z. (2004). *Some Nonparametric Regression Techniques for Complex Survey Data*. Unpublished Ph.D. thesis, The University of Western Ontario, London, Ontario, Canada.
- Zheng, H., and Little, R.J.A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, 30, 209-218.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2009.

R. Andridge, *Ohio State University*
 J.-F. Beaumont, *Statistics Canada*
 E. Berg, *Iowa State University*
 J.M. Brick, *Westat, Inc.*
 D. Cantor, *Westat Inc.*
 P. Cantwell, *U.S. Bureau of the Census*
 R. Chambers, *University of Wollongong, Australia*
 D. Chapman, *Federal Deposit Insurance Corporation*
 A.-S. Charest, *Carnegie-Mellon University*
 S. Chatterjee, *University of Minnesota*
 M. Cohen, *National Academy of Sciences/Committee on National Statistics*
 S. Cohen, *National Science Foundation*
 M.P. Couper, *University of Michigan*
 R. Curtin, *National Centre for Health Statistics*
 E. Dagum, *University of Bologna*
 G. Datta, *University of Georgia*
 P.-P. de Wolf, *Statistics Netherlands*
 P. Dick, *Statistics Canada*
 J. Dixon, *Bureau of Labor Statistics*
 J.L. Eltinge, *U.S. Bureau of Labor Statistics*
 V. Esteveao, *Statistics Canada*
 E. Fabrizi, *University of Bergamo, Italy*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistics Canada*
 M. Ghosh, *University of Florida*
 S. Godbout, *Statistics Canada*
 C. Goga, *Université de Bourgogne*
 B. Gross, *ABS*
 R.M. Groves, *U.S. Census Bureau*
 R. Harter, *National Opinion Research Centre*
 S. Haslett, *Massey University, New Zealand*
 D. Haziza, *Université de Montréal*
 M.A. Hidioglou, *Statistics Canada*
 G. James, *Office for National Statistics, UK*
 L. Jang, *Statistics Canada*
 J. Jiang, *University of California, Davis*
 D. Judkins, *Westat Inc.*
 C. Julien, *Statistics Canada*
 D. Kasprzyk, *Mathematica Policy Research*
 R.S. Kenett, *KPA Ltd., Raanana, Israel and University of Torino, Italy*
 J.-K. Kim, *Iowa State University*
 J.-M. Kim, *University of Minnesota-Morris*
 P. Kocio, *CSIRO*
 P. Kott, *National Agricultural Statistics Service*
 S. Laaksonen, *University of Helsinki*
 D. Ladiray, *INSEE*
 P. Lahiri, *JPSM, University of Maryland*
 P. Lavallée, *Statistics Canada*
 C. Leon, *Statistics Canada*
 R. Little, *University of Michigan*
 B. Liu, *Westat Inc.*
 L. Mach, *Statistics Canada*
 T. Maiti, *Iowa State University*
 H. Mantel, *Statistics Canada*
 J. Maples, *U.S. Census Bureau*
 A. Matei, *Université de Neuchâtel, Suisse*
 C. McLaren, *Office for National Statistics, UK*
 Y. McNab, *UBC*
 F. Mecatti, *University of Milan-Bicocca, Italy*
 S.M. Miller, *Bureau of Labor Statistics*
 L. Mohadjer, *Westat Inc.*
 G.E. Montinari, *University of Perugia, Italy*
 F.A.S. Moura, *Universidade do Brasil-UFRJ*
 Y. Mpetsheni, *Statistics South Africa*
 G. Nathan, *Hebrew University*
 T. Nayak, *George Washington University*
 J. Opsomer, *Colorado State University*
 S.P. Paben, *Bureau of Labor Statistics*
 M. Park, *Korea University*
 Z. Patak, *Statistics Canada*
 D. Pfeffermann, *Hebrew University*
 N.G.N. Prasad, *University of Alberta*
 M. Pratesi, *Università di Pisa*
 L. Qualité, *Université de Neuchâtel*
 J.N.K. Rao, *Carleton University*
 T.J. Rao, *Indian Statistical Institute*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 S. Rubin-Bleuer, *Statistics Canada*
 A. Ruiz-Gazen, *Université des Sciences Sociales de Toulouse*
 H. Saigo, *Waseda University*
 N. Salvati, *Università di Pisa*
 C.-E. Sæmndal, *Université de Montréal*
 O. Sautory, *INSEE*
 N. Schenker, *National Center for Health Statistics*
 F.J. Scheuren, *National Opinion Research Center*
 G. Shapiro, *Independent consultant*
 N. Shlomo, *University of Southampton*
 D.B.N. Silva, *Office for National statistics, U.K.*
 P. do N. Silva, *University of Southampton*
 S. Sinha, *Carleton University*
 C.J. Skinner, *University of Southampton*
 E. Slud, *University of Maryland and US Census Bureau*
 E. Stasny, *Ohio State University*
 D. Steel, *University of Wollongong*
 L. Stokes, *Southern Methodist University*
 M. Thompson, *University of Waterloo*
 Y. Tillé, *Université de Neuchâtel*
 R. Vaillant, *University of Maryland*
 V.J. Verma, *Università degli Studi di Siena*
 C. Walker, *Statistics Canada*
 D. Willimack, *U.S. Census Bureau*
 K.M. Wolter, *Iowa State University*
 C. Wu, *University of Waterloo*
 W. Yung, *Statistics Canada*
 A. Zaslavsky, *Harvard Medical School*
 F. Zhang, *National Science Foundation*

Acknowledgements are also due to those who assisted during the production of the 2009 issues: Eric Rancourt of Corporate Planning and Evaluation Division, Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau of Household Survey Methods Division, Nick Budko and Carole Jean-Marie of Business Survey Methods Division, Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Denis Coutu, Darquise Pellerin and Isabelle Poliquin (Dissemination Division), Sheri Buck (Systems Development Division) and Sylvie Dupont (Official Languages and Translation Division).

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 25, No. 2, 2009

Control Charts as a Tool for Data Quality Control Carl E. Pierchala, Jyoti Surti.....	167
Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates Emilia Peytcheva, Robert M. Groves	193
Analyzing Contact Sequences in Call Record Data. Potential and Limitations of Sequence Indicators for Nonresponse Adjustments in the European Social Survey Frauke Kreuter, Ulrich Kohler	203
Design and Estimation for Split Questionnaire Surveys James O. Chipperfield, David G. Steel	227
Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models Patrick Graham, Jim Young, Richard Penny	245
The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment Nobuaki Hoshino.....	269
Book and Software Reviews	293

Contents
Volume 25, No. 3, 2009

The Presentation of a Web Survey, Nonresponse and Measurement Error among Members of Web Panel Roger Tourangeau, Robert M. Groves, Courtney Kennedy, Ting Yan.....	299
Cooperation in Centralised CATI Household Panel Surveys – A Contact-based Multilevel Analysis to Examine Interviewer, Respondent, and Fieldwork Process Effects Oliver Lipps	323
Seam Effects in Quantitative Responses Frederick G. Conrad, Lance J. Rips, Scott S. Fricker	339
Testing a Cue-list to Aid Attitude Recall in Surveys: A Field Experiment Wander van der Vaart	363
Multipurpose Weighting for Small Area Estimation Hukum Chandra, Ray Chambers	379
A Note on the Effect of Auxiliary Information on the Variance of Cluster Sampling Nina Hagesæther, Li-Chun Zhang	397
Beyond Objective Priors for the Bayesian Bootstrap Analysis of Survey Data Cinzia Carota	405
Modeling Stock Trading Day Effects Under Flow Day-of-Week Effect Constraints David F. Findley, Brian C. Monsell	415

All inquires about submissions and subscriptions should be directed to jos@scb.se

Volume 37, No. 1, March/mars 2009

Paul GUSTAFSON Editor's report/Rapport du Rédacteur en chef	1
Ehab F. ABD-ELFATTAH & Ronald W. BUTLER Log-rank permutation tests for trend: saddlepoint p -values and survival rate confidence intervals	5
Imad BOU-HAMAD, Denis LAROCQUE, Hatem BEN-AMEUR, Louise C. MÂSSE, Frank VITARO & Richard E. TREMBLAY Discrete-time survival trees	17
Jerry BRUNNER & Peter C. AUSTIN Inflation of type I error rate in multiple regression when independent variables are measured with error	33
Jesse FREY An exact multinomial test for equivalence	47
Timothy HANSON, Wesley JOHNSON & Purushottam LAUD Semiparametric inference for survival models with step process covariates	60
Mhammed MESFIQUI, Jean-François QUESSY & Marie-Hélène TOUPIN On a new goodness-of-fit process for families of copulas	80
Xiao WANG Nonparametric estimation of the shape function in a gamma process for degradation data	102
Chunming ZHANG, Yuan JIANG & Zuofeng SHANG New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation	119
Acknowledgement of referees' services/Remerciements aux membres des jurys	140
Volume 37 (2009): Subscription rates/Frais d'abonnement	141

Volume 37, No. 2, June/juin 2009

Tim B. SWARTZ, Paramjit S. GILL & Saman MUTHUKUMARANA Modelling and simulation for one-day cricket	143
D.A.S. FRASER, A. WONG & Y. SUN Three enigmatic examples and inference from likelihood	161
Baojiang CHEN, Grace Y. YI & Richard J. COOK Likelihood analysis of joint marginal and conditional models for longitudinal categorical data	182
Vittorio ADDONA, Masoud ASGHARIAN & David B. WOLFSON On the incidence-prevalence relation and length-biased sampling	206
Sanjoy K. SINHA Bootstrap tests for variance components in generalized linear mixed models	219
Liang PENG A practical method for analysing heavy tailed data	235
José E. CHACÓN Data-driven choice of the smoothing parametrization for kernel density estimators	249
Peng ZHANG, Zhenguo QIU, Yuejiao FU & Peter X.-K. SONG Robust transformation mixed-effects models for longitudinal continuous proportional data	266
Xu ZHENG Testing heteroscedasticity in nonlinear and nonparametric regressions	282
Sujit K. SAHU, Dipak K. DEY & Márcia D. BRANCO <i>Erratum</i> : A new class of multivariate skew distributions with applications to Bayesian regression models	301
Volume 37 (2009): Subscription rates/Frais d'abonnement	303

Volume 37, No. 3, September/septembre 2009

Gail IVANOFF, Associate Editor, CJS In memory of André Robert Dabrowski	305
Herold DEHLING André Dabrowski's work on limit theorems and weak dependence	307
André DABROWSKI, Jiyeon LEE & David R. McDONALD Large deviations of multiclass $M/G/1$ queues	327
André DABROWSKI, Gail IVANOFF & Rafał KULIK Some notes on Poisson limits for empirical point processes	347
Raphael GOTTARDO & Adrian RAFTERY Bayesian robust transformation and variable selection: a unified approach	361
Sanjoy K. SINHA & J.N.K. RAO Robust small area estimation	381
Jean-François BEAUMONT & Cynthia BOCCI Variance estimation when donor imputation is used to fill in missing values	400
Hongmei ZHANG Designing sampling plans to capture rare objects	417
Lang WU, Wei LIU & Juxin LIU A longitudinal study of children's aggressive behaviours based on multivariate mixed models with incomplete data	435
Lieven DESMET & Irène GIJBELS Local linear fitting and improved estimation near peaks	453
Jaechoul LEE & Kyungduk KO First-order bias correction for fractionally integrated time series	476
Volume 37 (2009): Subscription rates/Frais d'abonnement	494

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
 - 1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(.) et log(.) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, I).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4. **Figures et tableaux**

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6. **Communications brèves**

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

Gail IVANOFF, Associate Editor, CJS	305
In memory of André Robert Dabrowski	
Herold DEHLING	307
André Dabrowski's work on limit theorems and weak dependence	
André DABROWSKI, Jiyoon LEE & David R. McDONALD	327
Large deviations of multiclass $M/G/1$ queues	
André DABROWSKI, Gail IVANOFF & Rafal KULIK	347
Some notes on Poisson limits for empirical point processes	
Raphael GOTTARDO & Adrian RAFTERY	361
Bayesian robust transformation and variable selection: a unified approach	
Sanjoy K. SINHA & J.N.K. RAO	381
Robust small area estimation	
Jean-François BEAUMONT & Cynthia BOCCI	400
Variance estimation when donor imputation is used to fill in missing values	
Hongmei ZHANG	417
Designing sampling plans to capture rare objects	
Lang WU, Wei LIU & Juxin LIU	435
A longitudinal study of children's aggressive behaviours based on multivariate mixed models with incomplete data	
Lieven DESMET & Irène GJBELS	453
Local linear fitting and improved estimation near peaks	
Jaechool LEE & Kyungduk KO	476
First-order bias correction for fractionally integrated time series	
Volume 37 (2009): Subscription rates/Frais d'abonnement	494

Volume 37, No. 1, March/mars 2009

Paul GUSTAFSON	1
Editor's report/Rapport du Rédacteur en chef	1
Bhab F. ABD-ELFATTAH & Ronald W. BUTLER	5
Log-rank permutation tests for trend: saddlepoint p -values and survival rate confidence intervals	5
Imad BOU-HAMAD, Denis LAROCQUE, Hatem BEN-AMEUR, Louise C. MÄSSE, Frank VITARO & Richard E. TREMBLAY	17
Discrete-time survival trees	17
Jerry BRUNNER & Peter C. AUSTIN	33
Inflation of type I error rate in multiple regression when independent variables are measured with error	33
Jesse FREY	47
An exact multinomial test for equivalence	47
Timothy HANSON, Wesley JOHNSON & Punishottam LAUD	60
Semiparametric inference for survival models with step process covariates	60
Mhamed MESFIoui, Jean-François QUESSY & Martine-Hélène TOUPIN	80
On a new goodness-of-fit process for families of copulas	80
Xiao WANG	102
Nonparametric estimation of the shape function in a gamma process for degradation data	102
Chunming ZHANG, Yuan JIANG & Zuofeng SHANG	119
New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation	119
Acknowledgement of referees' services/Remerciements aux membres des jurys	140
Volume 37 (2009): Subscription rates/Frais d'abonnement	141

Volume 37, No. 2, June/juin 2009

Tim B. SWARTZ, Paramjit S. GILL & Saman MUTHUKUMARANA	143
Modelling and simulation for one-day cricket	143
D.A.S. FRAISER, A. WONG & Y. SUN	161
Three enigmatic examples and inference from likelihood	161
Baojiang CHEN, Grace Y. YI & Richard J. COOK	182
Likelihood analysis of joint marginal and conditional models for longitudinal categorical data	182
Vittorio ADDONA, Masoud ASGHARIAN & David B. WOLFSON	206
On the incidence-prevalence relation and length-biased sampling	206
Sanjoy K. SINHA	219
Bootstrap tests for variance components in generalized linear mixed models	219
Liang PENG	235
A practical method for analysing heavy tailed data	235
Jose E. CHACÓN	249
Data-driven choice of the smoothing parametrization for kernel density estimators	249
Peng ZHANG, Zhenqiao QIU, Yueliao FU & Peter X.-K. SONG	266
Robust transformation mixed-effects models for longitudinal continuous proportional data	266
Xu ZHENG	282
Testing heteroscedasticity in nonlinear and nonparametric regressions	282
Sujit K. SAHU, Dipak K. DEY & Mária D. BRANCO	301
<i>Error</i> : A new class of multivariate skew distributions with applications to Bayesian regression models	301
Volume 37 (2009): Subscription rates/Frais d'abonnement	303

Contents Volume 25, No. 3, 2009

The Presentation of a Web Survey, Nonresponse and Measurement Error among Members of Web Panel Roger Tourangeau, Robert M. Groves, Courtney Kennedy, Ting Yan.....	299
Cooperation in Centralised CATI Household Panel Surveys – A Contact-based Multilevel Analysis to Examine Interviewer, Respondent, and Fieldwork Process Effects Oliver Lipps.....	323
Seam Effects in Quantitative Responses Frederick G. Conrad, Lance J. Rips, Scott S. Fricker.....	339
Testing a Cue-list to Aid Attitude Recall in Surveys: A Field Experiment Wander van der Vaart.....	363
Multipurpose Weighing for Small Area Estimation Hukum Chandra, Ray Chambers.....	379
A Note on the Effect of Auxiliary Information on the Variance of Cluster Sampling Nima Hagesæther, Li-Chun Zhang.....	397
Beyond Objective Priors for the Bayesian Bootstrap Analysis of Survey Data Cinzia Carota.....	405
Modeling Stock Trading Day Effects Under Flow Day-of-Week Effect Constraints David F. Findley, Brian C. Monsell.....	415

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

Contents

Volume 25, No. 2, 2009

Control Charts as a Tool for Data Quality Control	Carl E. Pierchala, Jyoti Surti.....	167
Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates	Emilia Peytcheva, Robert M. Groves	193
Analyzing Contact Sequences in Call Record Data. Potential and Limitations of Sequence Indicators for Nonresponse Adjustments in the European Social Survey	Frauke Kreuter, Ulrich Kohler	203
Design and Estimation for Split Questionnaire Surveys	James O. Chipperfield, David G. Steel	227
Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models	Patrick Graham, Jim Young, Richard Penny	245
The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment	Nobuaki Hoshino.....	269
Book and Software Reviews		293

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2009.

- J. Maples, *U.S. Census Bureau*
 A. Maret, *Université de Neuchâtel, Suisse*
 C. McLaren, *Office for National Statistics, UK*
 Y. McNab, *UBC*
 F. McNeill, *University of Milan-Bicocca, Italy*
 S.M. Miller, *Bureau of Labor Statistics*
 L. Mohadjer, *Westat Inc.*
 G.E. Montinaro, *University of Perugia, Italy*
 F.A.S. Moura, *Universidade do Brasil-LF-RJ*
 Y. Mpisameni, *Statistics South Africa*
 G. Nathan, *Hebrew University*
 T. Nayak, *George Washington University*
 S.P. Pabon, *Bureau of Labor Statistics*
 J. Opsomer, *Colorado State University*
 R. Curtin, *National Centre for Health Statistics*
 M.P. Cooper, *University of Michigan*
 S. Cohen, *National Science Foundation*
 M. Cohen, *National Academy of Sciences/Committee on National Statistics*
 S. Charney, *University of Minnesota*
 A.-S. Charest, *Carnegie-Mellon University*
 R. Chapman, *Federal Deposit Insurance Corporation*
 R. Chambers, *University of Wollongong, Australia*
 P. Cantwell, *U.S. Bureau of the Census*
 D. Cantor, *Westat Inc.*
 J.M. Brick, *Westat Inc.*
 E. Berg, *Iowa State University*
 J.-F. Beaumont, *Statistique Canada*
 R. Andridge, *Ohio State University*
 J.-F. Beaumont, *Statistique Canada*
 W.A. Fuller, *Iowa State University*
 E. Fabrizi, *University of Bergamo, Italy*
 V. Esteve, *Statistique Canada*
 R. Harter, *National Opinion Research Centre*
 S. Haslett, *Massy University, New Zealand*
 D. Haziza, *Université de Montréal*
 M.A. Hidiroglou, *Statistique Canada*
 G. James, *Office for National Statistics, UK*
 L. Jiang, *Statistique Canada*
 D. Juddkins, *Westat Inc.*
 C. Julien, *Statistique Canada*
 D. Kasprzyk, *Mathematica Policy Research*
 R.S. Kent, *KPA Ltd., Raanana, Israel and University of Torino, Italy*
 J.-M. Kim, *University of Minnesota-Morris*
 J.-K. Kim, *Iowa State University*
 P. Kokic, *CSIRO*
 P. Kott, *National Agricultural Statistics Service*
 S. Laaksonen, *University of Helsinki*
 D. Laditray, *INSEE*
 P. Lahiri, *JPSM, University of Maryland*
 P. Lavallée, *Statistique Canada*
 C. Leon, *Statistique Canada*
 R. Little, *University of Michigan*
 B. Liu, *Westat Inc.*
 L. Mach, *Statistique Canada*
 T. Mailli, *Iowa State University*
 H. Manel, *Statistique Canada*
 F. Zhang, *National Science Foundation*
 A. Zaslavsky, *Harvard Medical School*
 W. Yung, *Statistique Canada*
 K.M. Wolter, *Iowa State University*
 D. Williamson, *U.S. Census Bureau*
 C. Walker, *Statistique Canada*
 V.J. Verma, *Università degli Studi di Siena*
 R. Vaillan, *University of Maryland*
 Y. Tillie, *Université de Neuchâtel*
 M. Thompson, *University of Waterloo*
 D. Steel, *University of Wollongong*
 L. Stokes, *Southern Methodist University*
 E. Stasny, *Ohio State University*
 E. Stud, *University of Maryland and U.S. Census Bureau*
 C.J. Skinner, *University of Southampton*
 S. Sinha, *Carleton University*
 P. do N. Silva, *University of Southampton*
 D.B.N. Silva, *Office for National Statistics, UK*
 N. Shlomo, *University of Southampton*
 G. Shapito, *Independent consultant*
 F.J. Schewen, *National Opinion Research Center*
 O. Sautory, *INSEE*
 C.-E. Sändal, *Université de Montréal*
 H. Saigo, *Waseda University*
 N. Salvati, *Università di Pisa*
 A. Ruiz-Gazen, *Université des Sciences Sociales de Toulouse*
 A. Rubin-Bleuer, *Statistique Canada*
 L.-P. Rivest, *Université Laval*
 J. Reiter, *Duke University*
 T.J. Rao, *Indian Statistical Institute*
 J.N.K. Rao, *Carleton University*
 L. Quailie, *Université de Neuchâtel*
 M. Praest, *Università di Pisa*
 N.G.N. Prasad, *University of Alberta*
 D. Pfeffermann, *Hebrew University*
 Z. Patlak, *Statistique Canada*
 M. Park, *Korea University*
 S.P. Pabon, *Bureau of Labor Statistics*
 T. Nayak, *George Washington University*
 Y. Mpisameni, *Statistics South Africa*
 G. Nathan, *Hebrew University*
 T. Nayak, *George Washington University*
 S.P. Pabon, *Bureau of Labor Statistics*
 J. Opsomer, *Colorado State University*
 R. Curtin, *National Centre for Health Statistics*
 M.P. Cooper, *University of Michigan*
 S. Cohen, *National Science Foundation*
 M. Cohen, *National Academy of Sciences/Committee on National Statistics*
 S. Charney, *University of Minnesota*
 A.-S. Charest, *Carnegie-Mellon University*
 R. Chapman, *Federal Deposit Insurance Corporation*
 R. Chambers, *University of Wollongong, Australia*
 P. Cantwell, *U.S. Bureau of the Census*
 D. Cantor, *Westat Inc.*
 J.M. Brick, *Westat Inc.*
 E. Berg, *Iowa State University*
 J.-F. Beaumont, *Statistique Canada*
 R. Andridge, *Ohio State University*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2009 : Eric Rancourt de la Division de la planification intégrée et de l'évaluation, Céline Ethier de la Division de la recherche et de l'innovation en statistique, Christine Cousineau de la Division des méthodes d'enquêtes auprès des ménages, Nick Budko et Carole Jean-Marie de la Division des méthodes d'enquêtes auprès des entreprises, Cécile Bouquet, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Denis Cout, Darquise Pellerin et Isabelle Poliquin (Division de la diffusion), Sheri Buck (Division du développement des systèmes) et Sylvie Dupont (Division des langues officielles et traduction).

les résultats de l'« astuce de Cramér-Wold » (*Cramér-Wold device*) donnés dans Serfling (1980, page 18), nous voyons que $\sqrt{n}(\bar{y} - \bar{x}B - \bar{R})$ converge vers un vecteur aléatoire normalement distribué. Donc, en utilisant cette idée indirecte de Slutsky, nous avons prouvé la normalité de $\sqrt{n}(\bar{r} - \bar{R}) = \sqrt{n}(\bar{y} - \bar{x}B - \bar{R})$.

A.5 Preuve du théorème 2

La preuve découle du même argument que dans la preuve du théorème 2.2.1 de Wang (2004), selon lequel $\hat{g}(z_g)$ est une fonction de la moyenne et des proportions de domaine.

Remerciements

Les présents travaux ont été financés par une subvention du Conseil de recherches en sciences naturelles et en génie (CRSNG) du Canada. L'auteur remercie Mary Thompson de ses suggestions et commentaires constructifs formulés lors d'une version antérieure du présent article. Les auteurs remercient aussi le rédacteur associé et les deux examinateurs de leurs commentaires très utiles.

Bibliographie

Belhoucne, D.R., et Stafford, J.E. (1999). Density estimation from complex survey. *Statistica Sinica*, 9, 407-424.

Belhoucne, D.R., et Stafford, J.E. (2001). Régression polynomiale locale dans le cas des enquêtes complexes. *Techniques d'enquête*, 27, 219-226.

Bickel, P.J., et Freedman, D.A. (1983). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

Buskirk, D.T., et Lohr, L.S. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Fan, J., et Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society, Séries B*, 57, 371-394.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C, 37, 117-132.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudományos Akadémia Budapest Matematikai Kísérleti Kutató Intézet Közleményei*, 5, 361-374.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Jones, M.C. (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84, 733-741.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balance repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535-545.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.

Ontario Health Survey (1992). *Ontario Health Survey: User's Guide*. Ministry of Health, Toronto, Ontario, Canada.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2^{ème} Ed.). New York : John Wiley & Sons, Inc.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56, 931-954.

Sändal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Serfling, R.J. (1980). *Approximation Theorem of Mathematical Statistics*. New York : John Wiley & Sons, Inc.

Shao, J. (1996). Resampling methods in sample survey. *Statistics*, 27, 203-254.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Séries B*, 50, 413-436.

Thompson, M.E. (1997). *Theory of Sample Survey* (1^{ère} Ed.). New York : Chapman and Hall.

Wang, Z. (2004). *Some Nonparametric Regression Techniques for Complex Survey Data*. Thèse de doctorat non-publiée, The University of Western Ontario, Londres, Ontario, Canada.

Zheng, H., et Little, R.J.A. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 233-243.

$$\sqrt{n}/N \|\hat{\theta} - \theta\|_p \rightarrow 0.$$

L'argument standard présenté dans Rao (1973, page 387)

$$+ \sqrt{n} \|\hat{\theta} - \theta\|_2.$$

$$- \frac{\sqrt{n} \hat{u}_B(\theta)}{N} (\hat{B} - B) = \left(\frac{\sqrt{n}}{N} (u(\theta) - u(\theta)) + U_{\xi}(\theta) \frac{\sqrt{n}}{N} (m_{\xi}(z) - m_{\xi}(z)) \right)$$

Puisque, pour le vrai θ , nous avons $u(\theta) = 0$, nous pouvons réécrire l'équation (13) comme il suit :

A.3 Preuve du théorème 1

Chaque entrée de $\hat{m}_{\xi}(z)$ est simplement une fonction de régression estimée par la méthode des polynômes locaux établie par Bellhouse et Stafford (2001). Le théorème 2.2.1 dans Wang (2004) montre que $\hat{m}_{\xi}(z)$ est convergent à la vitesse racine carrée de n . Donc, puisque la dimension de $\hat{m}_{\xi}(z)$ est finie, nous pouvons montrer que $\sqrt{n}(\hat{m}_{\xi}(z) - m_{\xi}(z))$ est borné en probabilité.

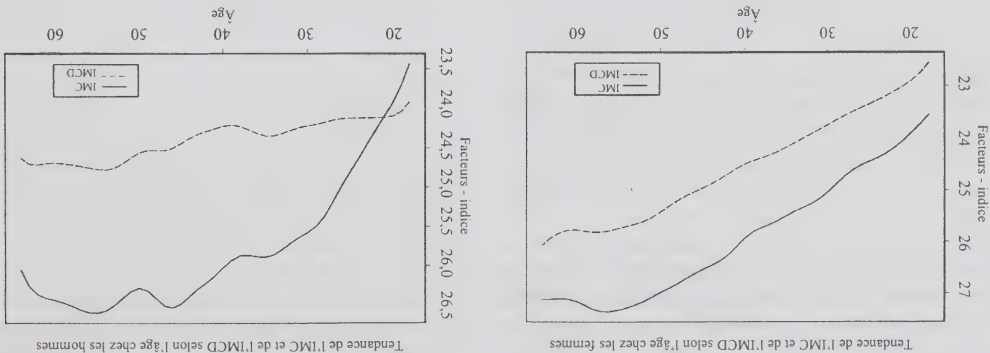
A.2 Preuve du lemme 2

En observant que les entrées de $\hat{u}(\theta)$, $\hat{u}_B(\theta)$ et $U_{\xi}(\theta)$ sont soit des totaux d'échantillon ou des ratios de totaux d'échantillon, nous pouvons appliquer les lemmes 1.2.5 et 1.2.6 présentés dans Wang (2004) pour établir ce lemme.

A.1 Preuve du lemme 1

Annexe

Figure 5 Comparaison des tendances estimées de l'IMC et de l'IMCD selon l'âge chez les hommes qui sont actifs ou moyennement actifs



Le tableau 3 donne les estimations par sondage des coefficients linéaires dans les modèles (a) et (b). Si nous comparons l'IMC selon le sexe, nous constatons qu'il est plus élevé chez les hommes. Si la catégorie de base est ancien fumeur, les coefficients de la situation d'usage du tabac sont tous négatifs et significatifs, ce qui fait penser que les anciens fumeurs ont tendance à être plus lourds que les personnes qui ont une autre situation d'usage du tabac. Les estimations indiquent aussi que les personnes inactives ont un IMC plus élevé. En ce qui concerne l'IMCD, les valeurs p suggèrent que la plupart des facteurs liés au mode de vie n'ont pas d'effet significatif.

Tableau 3
Résultats pour les modèles de régression semiparamétrique (a) et (b) (les erreurs-types sont indiquées entre parenthèses)

Variable	B_1	valeur p	B_2	valeur p
Sexe	1,45 (0,05)	0,00	2,80 (0,05)	0,00
N'a jamais fumé	-0,45 (0,10)	0,00	-0,06 (0,06)	0,34
Fumée à l'occasion	-0,31 (0,17)	0,04	-0,00 (0,10)	0,96
Fumée quotidiennement	-0,61 (0,09)	0,00	-0,12 (0,06)	0,03
Moyennement actif	-0,33 (0,09)	0,00	-0,07 (0,06)	0,24
Actif	-0,50 (0,09)	0,00	-0,14 (0,06)	0,07

À la figure 4, les fonctions estimées de l'âge, $\hat{g}_1(\text{Âge})$ et $\hat{g}_2(\text{Âge})$, et leurs bandes de confiance sont représentées graphiquement en fonction de l'âge. Nous constatons que, dans les deux cas, l'IMC et l'IMCD sont des fonctions croissantes de l'âge.

La figure 5 donne les fonctions estimées de l'âge, $\hat{g}_1(\text{Âge})$ et $\hat{g}_2(\text{Âge})$, pour les personnes actives et moyennement actives. Si nous examinons l'effet de l'âge séparément chez les femmes et les hommes, nous voyons que chez les femmes actives ou moyennement actives, l'IMCD est, en moyenne, plus faible que l'IMC, tandis que les hommes dont l'intensité d'activité physique est la même manifestent le désir d'augmenter leur poids avant l'âge de

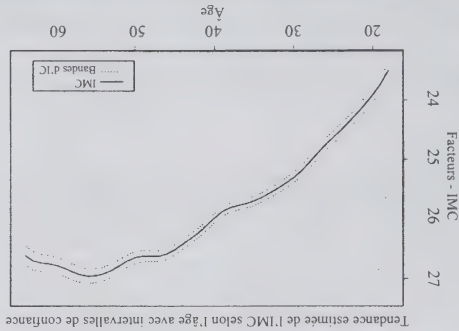
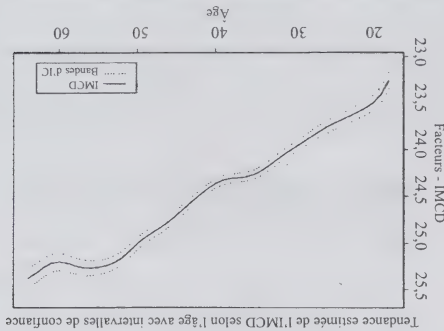


Figure 4 Tendances estimées de l'IMC et de l'IMCD selon l'âge avec intervalles de confiance à 95 % ponctuels



Dans le modèle linéaire partiel de travail, nous supposons qu'il n'existe aucune interaction entre la composante paramétrique et la variance des coefficients linéaires et de la variance des termes d'interaction consiste à laisser la composante non paramétrique paraître linéaire dans le modèle d'interaction. Autrement dit, nous définissons le modèle linéaire partiel sous la forme

$$y = G(z) + X\beta + XH(z) + \epsilon.$$

En testant l'écart de $H(z)$ par rapport à zéro, nous pouvons détecter l'existence d'une interaction.

Pour estimer l'espérance conditionnelle sur les composantes non paramétriques pour des variables indicatrices aléatoires discrètes, nous proposons d'utiliser des modèles linéaires ou additifs généralisés.

6. Conclusion

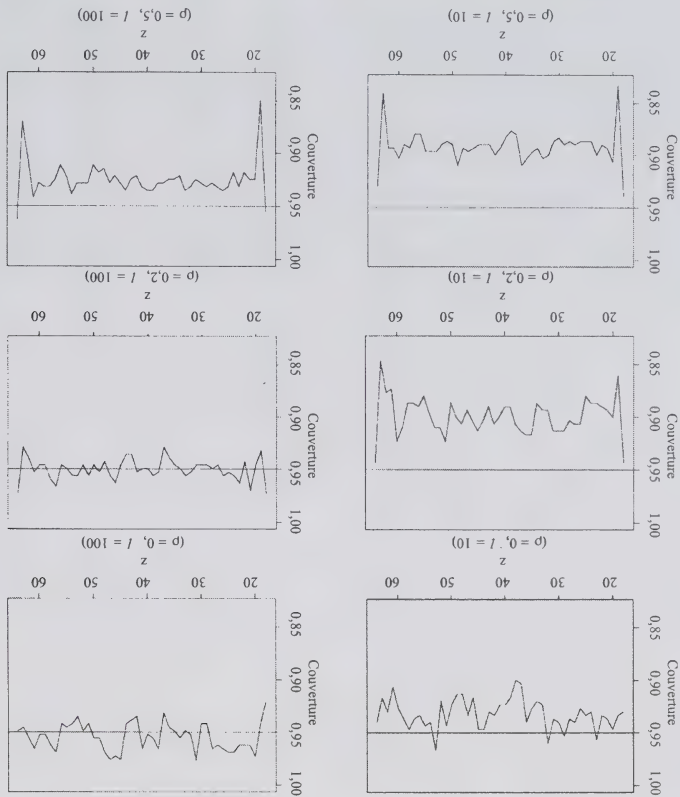
Àu moyen d'un modèle linéaire partiel, nous étendons les méthodes de régression semiparamétrique aux données d'enquêtes complexes. Nous élaborons les propriétés asymptotiques des estimateurs par sondage. Le calcul des estimations de la variance des coefficients linéaires et de la fonction de régression s'appuie sur les estimations de variance des totaux et des moyennes d'enquête. À condition d'obtenir les estimations de variance requises de ces totaux et moyennes, nous pouvons appliquer la méthode en nous servant de logiciels statistiques standard.

Dans le modèle linéaire partiel de travail, nous supposons qu'il n'existe aucune interaction entre la composante paramétrique et la variance des coefficients linéaires et de la variance des termes d'interaction consiste à laisser la composante non paramétrique paraître linéaire dans le modèle d'interaction. Autrement dit, nous définissons le modèle linéaire partiel sous la forme

Dans le modèle linéaire partiel de travail, nous supposons qu'il n'existe aucune interaction entre la composante paramétrique et la variance des coefficients linéaires et de la variance des termes d'interaction consiste à laisser la composante non paramétrique paraître linéaire dans le modèle d'interaction. Autrement dit, nous définissons le modèle linéaire partiel sous la forme

En testant l'écart de $H(z)$ par rapport à zéro, nous pouvons détecter l'existence d'une interaction.

Pour estimer l'espérance conditionnelle sur les composantes non paramétriques pour des variables indicatrices aléatoires discrètes, nous proposons d'utiliser des modèles linéaires ou additifs généralisés.

Figure 3 Couverture des intervalles de confiance à 95 % ponctuels pour $g(z)$ 

5. Exemples empiriques

Revenons maintenant à l'exemple de la section 1. Afin d'illustrer le modèle linéaire partiel, nous examinons les effets de l'âge, du sexe, de la situation d'usage du tabac et de l'activité physique sur l'indice de masse corporelle (IMC) et sur l'indice de masse corporelle désire (IMCD). Construite comme l'IMC, l'IMCD est une variable dérivée produite d'après les réponses à la question sur le poids

Soulignons que, si la taille des unités primaires d'échantillonnage est grande (1 000), la fraction d'échantillonnage est très faible (0,05). Par conséquent, ces propriétés des estimations ne changeraient pas même si la taille des unités primaires d'échantillonnage était faible.

plus élevée que le niveau nominal, parce que l'effet de borne de l'estimation par la régression polynomiale locale cause un plus grand biais aux deux bornes des données. Pour $p = 0,5$, la taille effective d'échantillon est faible, de sorte que l'effet de borne devient sévère, ce qui crée les pics dirigés vers le bas à 18 et à 63.

souhaité par la personne. Puisque la croissance des personnes faisant partie du groupe d'âge qui nous intéresse est terminée, nous nous servons de la taille réelle pour calculer l'IMCD. Nous utilisons l'âge comme covariable non paramétrique et traitons les autres facteurs comme des variables discrètes. Comme il n'existe que 47 points distincts dans la variable d'âge, nous groupons l'ensemble de données par classe en fonction de l'âge. La taille de classe est fixée à l'unité, de sorte qu'il existe 47 classes, dont les points médians sont 18, 19, ..., 64. Parmi les variables explicatives catégoriques, le sexe possède deux niveaux : masculin = 1 et féminin = 0 ; la situation d'usage du tabac comprend les niveaux tels que ancien fumeur = 0, n'a jamais fumé = 1, fume à l'occasion = 2, fume quotidiennement = 3 ; et l'activité physique est répartie en trois niveaux : personne active = 1 et personne moyennement active = 0, personne inactive = 2. Les modèles de régression sont a) $IMC = g_1(\text{âge}) + \mathbf{XB}_1 + \mathbf{e}_1$ et b) $IMCD = g_2(\text{âge}) + \mathbf{XB}_2 + \mathbf{e}_2$, où \mathbf{X} est la matrice de plan contenant toutes les variables indicatrices.

Tableau 1
Résultats des simulations pour les estimateurs ponctuels de B

l	B_1	B_2	$p = 0$				$p = 0,2$				$p = 0,5$			
			Var ($\times 10^{-3}$)	Biais ² ($\times 10^{-6}$)	Rvar	Biais ² ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Biais ² ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Var ($\times 10^{-3}$)	Rvar	Biais ² ($\times 10^{-6}$)	Var ($\times 10^{-3}$)
10	5,77	1,07	0,46	0,21	0,13	3,12	1,1	1,01	0,23	1,19	1,33	0,98	1,19	1,33
25	9,97	0,97	0,54	0,21	1,08	0,38	0,44	1,08	0,30	0,53	0,98	0,98	0,53	0,98
50	0,22	0,13	0,22	0,13	1,07	0,13	0,27	0,93	0,026	0,21	1,18	0,98	0,21	1,18
100	0,36	0,13	0,36	0,13	0,96	0,11	1,06	0,039	0,13	0,13	1,78	0,98	0,13	1,78
10	3,32	1,31	0,64	0,31	1,10	1,54	1,34	0,92	1,26	3,5	1,78	0,98	1,26	3,5
25	0,75	0,75	0,31	0,31	1,07	2,40	1,06	1,42	0,14	1,42	1,03	0,97	0,14	1,42
50	0,38	0,94	0,15	0,15	0,94	0,85	0,94	0,16	0,072	0,33	1,03	0,97	0,16	0,76
100	0,38	0,94	0,15	0,15	0,94	0,38	0,98	0,072	0,33	0,33	1,03	0,97	0,33	1,03

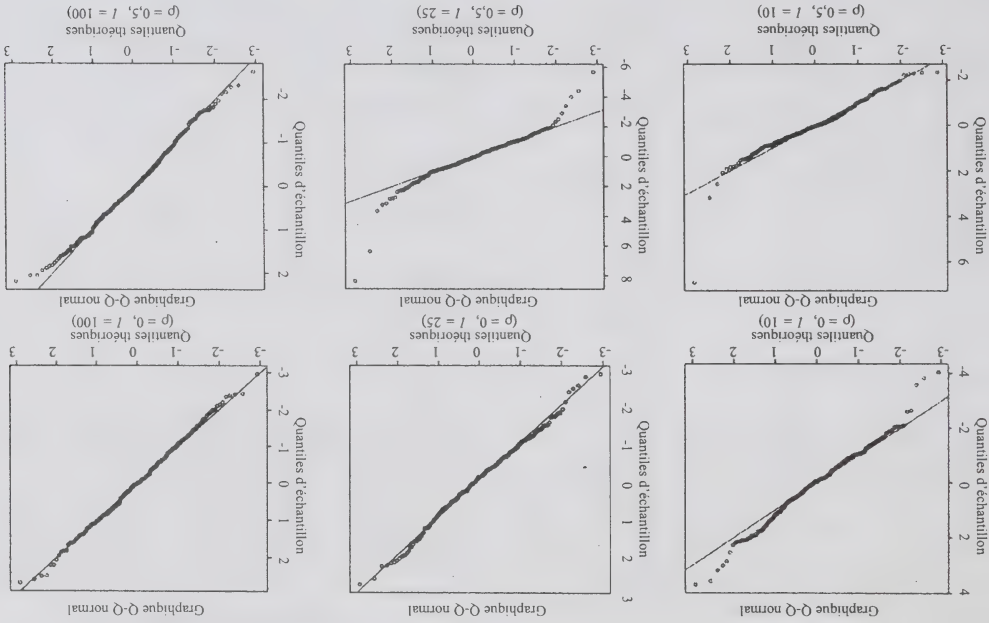


Figure 2 Graphiques quantile-quantile pour B_2 standardisé

Tableau 2
Biais et efficacité de $g(z)$

p	l	MEQMI	EQMI	MBiais ² ($\times 10^{-5}$)	Rvar
0	10	0,37	0,42	5,29	1,27
	25	0,15	0,17	3,20	1,10
	50	0,074	0,086	3,29	1,09
	100	0,037	0,044	2,34	1,08
0,2	10	2,95	3,25	6,13	0,91
	25	1,22	1,17	3,71	1,04
	50	0,74	0,54	2,34	1,0
	100	0,26	0,27	7,08	0,98
0,5	10	8,143	8,877	3,73	0,92
	25	3,155	3,073	6,56	1,03
	50	1,461	1,599	2,86	1,15
	100	0,659	0,607	3,59	1,09

À la figure 3, la couverture des intervalles de confiance à 95 % ponctuel pour $g(\cdot)$ varie entre 85 % et 96 %. Elle s'améliore à mesure qu'augmente la taille de l'échantillon. Toutefois, la performance de $g(\cdot)$ est plus sensible à la réduction de la taille effective d'échantillon causée par la corrélation intragroupe. En particulier, les couvertures des intervalles de confiance à 95 % dans les cas où $p = 0,2$ et $p = 0,5$ sont inférieures au niveau de confiance nominal de 95 % quand $l = 10$. La couverture s'améliore à mesure que le nombre d'unités primaires d'échantillonnage augmente pour les cas où $p = 0$ et $p = 0,2$. Pour $p = 0,5$, le défaut de couverture persiste quand la taille d'échantillon croît jusqu'à 100. Nous voyons aussi qu'à $z = 18$ ou 64, la couverture est

4.2 Résultats

présentons que les cas où $l = 10, 25, 100$ et $p = 0, 0,5$. La performance globale des estimateurs concorde avec la théorie des théorèmes 1 et 2.

Le tableau 1 confirme l'absence de biais par rapport au plan dans **B**. Il montre aussi que, à mesure que la taille d'échantillon augmente, les propriétés des estimations des coefficients linéaires s'améliorent pour toutes les structures d'erreur. En particulier, le carré du biais et la variance de **B** diminuent quand le nombre d'échantillons primaires augmente. La variance estimée de **B** se rapproche de la variance simulée de **B** à mesure que la taille d'échantillon augmente, ce qui confirme la convergence des estimations de variance de **B**. Si nous comparons les variances et les biais de **B** dans les cas où $p = 0,2$ et $p = 0,5$, au cas où $p = 0$, nous constatons que la corrélation intragrappe (effet de grappe) n'a pas eu d'incidence sur les propriétés de **B**. Cela pourrait être dû au fait que la taille d'échantillon intragrappe était grande.

L'examen de la figure 2 révèle que le nombre d'unités primaires échantillonnées et l'effet de grappe jouent un certain rôle dans la normalité de **B**. En particulier, quand la taille de l'échantillon est faible, par exemple $l = 10$, la normalité de l'estimateur **B** standardisé présente un certain écart par rapport à la théorie pour $p = 0$ et $p = 0,5$. Quand l passe à 25, nous constatons que la performance de **B** pour $p = 0$ commence à s'améliorer, tandis que pour $p = 0,5$, nous n'observons aucune amélioration jusqu'à $l = 100$. Empiriquement, ce résultat donne à penser que si le nombre de grappes est faible, nous ne devrions pas nous fier à la normalité théorique des estimations du coefficient; nous pourrions plutôt utiliser la loi t pour effectuer l'inférence.

En ce qui concerne les résultats de la partie non paramétrique de l'estimation, le tableau 2 montre que la moyenne des estimations de l'erreur quadratique moyenne intégrée simulee pour toutes les tailles d'échantillon et intégrées simulee des erreurs quadratiques moyennes est de nouveau confirmée par la moyenne du carré du biais (MBias²). Les valeurs du ratio moyen de la variance estimée à la variance simulée (RVar), qui sont proches de 1 dans tous les cas, sont en harmonie avec la convergence par rapport au plan de l'estimateur de $\hat{g}(z)$. Les erreurs quadratiques moyennes intégrées de $\hat{g}(\cdot)$ sont influencées par les corrélations intragrappe, ce que l'on peut déduire du fait que l'erreur quadratique moyenne intégrée ainsi que l'erreur quadratique moyenne intégrée estimée comportent de manière semblable à ceux de B_1 . La figure 3 souligne que les graphiques quantile-quantile de B_1 se comportent de manière semblable à ceux de B_2 . La figure 3 représente graphiquement la couverture des intervalles de confiance à 95 % pour $g(\cdot)$. Dans les figures 2 et 3, nous ne

En utilisant la population finie produite, nous avons trouvé que les estimations sous recensement étaient $B_1 = 2,01$ et $B_2 = 3,00$. Pour confirmer l'absence de biais par rapport au plan et l'efficacité de **B**, nous avons calculé le carré du biais simulé (Biais²), qui est le carré de la différence entre la moyenne des estimations simulées et des estimations sous recensement. En outre, nous présentons le ratio des estimations moyennes de la variance à la variance simulée de chaque estimateur d'un coefficient linéaire (RVar) pour montrer la validité de l'estimateur de variance standardisé B_1 . Pour évaluer la normalité de **B**, nous avons utilisé l'écart-type empirique et la valeur de population de **B**, et nous avons tracé les graphiques quantile - quantile des valeurs standardisées.

En appliquant la méthode semiparamétrique de Speckman (1988) au modèle (17), nous avons obtenu les estimations sous recensement $\hat{g}(z)$ pour $z = 18, \dots, 64$. Pour évaluer l'exactitude de $\hat{g}(z)$ par rapport au plan, nous avons calculé la différence entre $\hat{g}(z)$ et $g(z)$ à chaque point distinct. La moyenne des carrés des différences sur 47 valeurs distinctes de z est alors présentée comme le carré du biais moyen MBias². Nous avons calculé deux erreurs quadratiques moyennes pour vérifier l'efficacité par rapport au plan de $\hat{g}(z)$ et la convergence de $\text{Var}_p(\hat{g}(z))$. L'une d'elles est la moyenne des estimations de l'erreur quadratique moyenne intégrée (MEQM), que nous obtenons en calculant d'abord la somme des $\text{Var}_p(\hat{g}(z))$ sur $z = 18, \dots, 64$ pour chaque simulation, puis en prenant la moyenne des sommes sur le nombre total des simulations. L'erreur quadratique moyenne intégrée (EQMI) simulée est une autre erreur quadratique moyenne que nous avons calculé par sommation des erreurs quadratiques moyennes simulées à chaque point distinct de z . La moyenne des ratios de la moyenne simulée de $\text{Var}_p(\hat{g}(z))$ à la variance simulée de $\hat{g}(z)$ (Reff) montre la convergence de l'intervalle de confiance à 95 % ponctuel à chaque point distinct de z .

Les résultats pour les propriétés de **B**, $\text{Var}_p(\hat{g}(z))$ et $\text{Var}_p(\hat{g}(z))$ sont présentés aux tableaux 1 et 2 et aux figures 2 et 3. Les tableaux 1 et 2 donnent l'information sur l'exactitude et la précision des estimations simulées de **B** et $\hat{g}(\cdot)$. La figure 2 donne les graphiques quantile-quantile de la valeur normalisée en échantillon de B_2 . Il convient de souligner que les graphiques quantile-quantile de B_1 se comportent de manière semblable à ceux de B_2 . La figure 3 représente graphiquement la couverture des intervalles de confiance à 95 % pour $g(\cdot)$. Dans les figures 2 et 3, nous ne

4. Études par simulation

4.1 Plan de l'expérience

L'étude par simulation dont l'exécution est décrite ici a été conçue en vue d'illustrer les résultats théoriques des théorèmes 1 et 2. Nous avons produit les données suivant un processus en deux étapes qui imitait l'approche d'échantillonnage en superpopulation. Pour commencer, nous avons produit la population finie, puis nous avons sélectionné l'échantillon à partir de cette population. En particulier, nous avons considéré une population finie de $L = 500$ grappes avec $M (= M_j) = 20\,000$ dans chacune d'elles. Nous avons obtenu les observations de population pour la mesure d'intérêt y_{ij} à partir du modèle

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + 0,5 \exp\left(\frac{10}{z_{ij}} - 40\right) + \mu_i + \varepsilon_{ij} \quad (17)$$

pour $i = 1, \dots, L$ et $j = 1, \dots, M$ où les termes d'erreur μ_i et ε_{ij} sont mutuellement indépendants avec $\mu_i \sim N(0, \sigma_\mu^2)$ et $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Nous posons que $\sigma_\varepsilon^2 = \sigma_\mu^2 + \sigma_a^2$, de sorte que le coefficient de corrélation intragrappe est $\rho = \sigma_\mu^2 / \sigma_\varepsilon^2$. Parmi les covariables comprises dans le modèle, x_{1ij} ainsi que x_{2ij} ont été traitées comme étant la partie linéaire paramétrique du modèle et z_{ij} , comme étant la partie non paramétrique. Nous avons produit les x_{1ij} à partir de la loi de Bernoulli(1/2) et les x_{2ij} à partir de la distribution uniforme(0, 1). Les z_{ij} ont été produites (selon le butin de l'âge de la population canadienne (selon le Recensement de 1996) pour la tranche d'âge de 18 à 64 ans et elles étaient indépendantes des termes d'erreur. Nous présentons dans cette étude les résultats pour les valeurs $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$, $\sigma_\varepsilon^2 = 3$ et $\rho = 0,0,2,0,5$. Nous avons utilisé pour l'étude un plan d'échantillonnage à deux degrés, avec $l (= 10, 25, 50, 100)$ grappes sélectionnées au hasard dans L et $m (= 1\,000)$ unités secondaires d'échantillonage sélectionnées au hasard dans chaque grappe de taille M . Pour chaque taille d'échantillon et valeur de ρ , nous avons répété la simulation 300 fois. Au niveau de la population, nous avons appliqué la méthode de sélection de la largeur de fenêtre de Fan et Gijbels (1995) et déterminé que les largeurs de fenêtre pour l'estimation des espérances conditionnelles de X_1 et X_2 sur z étaient de 1,2 et 1,5 respectivement. Quand nous avons lissé les résidus pour estimer $g(z)$, la largeur de fenêtre était de 0,6.

$$\text{Var}_p(\mathbf{r}) = (\mathbf{Q} \otimes \mathbf{I}_D) \text{Cov}_p(\mathbf{x}, \mathbf{y}) (\mathbf{Q} \otimes \mathbf{I}_D)^T$$

$$+ \mathbf{x} \text{Var}_p(\mathbf{b}) \mathbf{x}^T$$

$$- 2(\mathbf{Q} \otimes \mathbf{I}_D) \text{Cov}_p(\mathbf{y}, \mathbf{b}) \mathbf{x}^T$$

$$- \sum_{j=1}^p 2(\mathbf{Q} \otimes \mathbf{I}_D) \text{Cov}_p(\mathbf{x}_j, \mathbf{b}) \mathbf{x}_j^T.$$

Sachant la variance estimée de \mathbf{r} , à savoir

$$\widehat{\text{Var}}_p(\mathbf{r}) = (\widehat{\mathbf{Q}} \otimes \mathbf{I}_D) \widehat{\text{Cov}}_p(\mathbf{x}, \mathbf{y}) (\widehat{\mathbf{Q}} \otimes \mathbf{I}_D)^T$$

$$+ \mathbf{x} \widehat{\text{Var}}_p(\mathbf{b}) \mathbf{x}^T$$

$$- 2(\widehat{\mathbf{Q}} \otimes \mathbf{I}_D) \widehat{\text{Cov}}_p(\mathbf{y}, \mathbf{b}) \mathbf{x}^T$$

$$- \sum_{j=1}^p 2(\widehat{\mathbf{Q}} \otimes \mathbf{I}_D) \widehat{\text{Cov}}_p(\mathbf{x}_j, \mathbf{b}) \mathbf{x}_j^T,$$

la variance estimée de $\widehat{g}(z_d)$ est $\widehat{\text{Var}}_p(\widehat{g}(z_d)) = \widehat{\mathbf{A}}_p^T \widehat{\text{Var}}_p(\mathbf{r}) \widehat{\mathbf{A}}_p$.

La normalité asymptotique de $\widehat{g}(\cdot)$ dépend aussi de la normalité de \mathbf{r} , qui est montrée dans le lemme suivant.

Lemme 3. Sous les conditions C1 à C7 et en supposant que la dimension de \mathbf{r} est finie, nous avons, quand v tend vers l'infini

$$\sqrt{n}(\mathbf{r} - \mathbf{R}) \xrightarrow{D} N(\mathbf{0}, \mathbf{V}(\mathbf{r})),$$

où

$$\mathbf{V}(\mathbf{r}) = \lim_{v \rightarrow \infty} n \text{Var}_p(\mathbf{r}).$$

En nous basant sur la normalité asymptotique établie dans le lemme 3 et sur l'estimateur de la variance de $\widehat{g}(z_d)$, nous établissons les propriétés asymptotiques de $\widehat{g}(z_d)$ dans le théorème suivant.

Théorème 2. Sous les conditions C1 à C7, nous avons, quand v tend vers l'infini :

$$\begin{aligned} & |\widehat{g}(z_d) - g(z_d)| \xrightarrow{P} 0; \\ & \sqrt{n}(\widehat{g}(z_d) - g(z_d)) \xrightarrow{D} N(0, 1). \end{aligned}$$

$$1) \sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) \rightarrow N(0, \mathbf{V}(\hat{\mathbf{B}})) \text{ où } \mathbf{V}(\hat{\mathbf{B}}) = \lim_{n \rightarrow \infty} n \text{Var}^p(\hat{\mathbf{B}});$$

$$2) \|\hat{\mathbf{B}} - \mathbf{B}\| \text{ converge vers zéro en probabilité.}$$

Pour obtenir les moments approximatifs de $\hat{\mathbf{B}}$, nous prenons les espérances des deux membres de l'équation (13), ce qui nous donne

$$E^p(-\hat{\mathbf{u}}_{\mathbf{B}}(\theta)(\hat{\mathbf{B}} - \mathbf{B})) = E^p(\hat{\mathbf{u}}(\theta))$$

$$+ E^p\{\mathbf{U}^{\xi}(\theta)[\hat{\mathbf{m}}^{\xi}(\mathbf{z}) - \mathbf{m}^{\xi}(\mathbf{z})]\}$$

$$+ E^p(\|\hat{\theta} - \theta\|) \mathbf{e}. \quad (15)$$

L'hypothèse selon laquelle les deuxièmes moments des estimations sont bornés fait disparaître le dernier terme de l'équation (15) à la limite. En nous inspirant de Binder (1983), nous avons

$$E^p(-\hat{\mathbf{u}}_{\mathbf{B}}(\theta))(E^p((\hat{\mathbf{B}} - \mathbf{B})) = E^p(\hat{\mathbf{u}}(\theta))$$

$$+ E^p(\mathbf{U}^{\xi}(\theta))E^p\{\hat{\mathbf{m}}^{\xi}(\mathbf{z}) - \mathbf{m}^{\xi}(\mathbf{z})\}.$$

Les totaux d'enquête qui définissent le vecteur $\hat{\mathbf{u}}(\theta)$ et la matrice $\hat{\mathbf{u}}_{\mathbf{B}}(\theta)$ sont des estimateurs de type Horvitz-Thompson et sont sans biais (Thompson 1997). D'où $E^p(\hat{\mathbf{u}}(\theta)) = \mathbf{u}(\theta)$ et $E^p(\hat{\mathbf{u}}_{\mathbf{B}}(\theta)) = \mathbf{u}_{\mathbf{B}}(\theta)$. Puisque $\mathbf{u}(\theta)$ est l'équation d'estimation des coefficients linéaires partiels définis en (8), elle équivaut à un vecteur nul de dimension $1 \times p$. En outre, Bellhouse et Stafford (2001) ont montré que $\hat{\mathbf{m}}^{\xi}(\mathbf{z})$ est un estimateur asymptotiquement sans biais de $\mathbf{m}^{\xi}(\mathbf{z})$. Donc, $-\mathbf{u}_{\mathbf{B}}(\theta)E^p(\hat{\mathbf{B}} - \mathbf{B}) = 0$, ou, en nous basant sur les conditions que $\mathbf{u}_{\mathbf{B}}(\theta)$ est inversible et que $\mathbf{u}_{\mathbf{B}}(\theta)^{-1}$ est finie, nous avons $E^p(\hat{\mathbf{B}}) = \mathbf{B}$.

En prenant la variance des deux membres de l'équation (13) et en utilisant les matrices de variance-covariance asymptotiques de $\hat{\mathbf{u}}(\theta)$ et $\hat{\mathbf{m}}^{\xi}(\mathbf{z})$, nous obtenons la variance

$$\text{Var}^p(\hat{\mathbf{B}}) = \mathbf{u}_{\mathbf{B}}(\theta)^{-1} \\ (\text{Var}^p(\hat{\mathbf{u}}(\theta)) + \mathbf{U}^{\xi}(\theta)(\mathbf{A}(\mathbf{I} \otimes \text{Cov}^p(\hat{\mathbf{x}}, \hat{\mathbf{y}}))\mathbf{A}^T) \mathbf{U}^{\xi}(\theta)^T \\ + 2(\mathbf{I}^p \otimes \ell)(\ell \otimes \mathbf{C})\mathbf{A}^T \mathbf{U}^{\xi}(\theta)(\mathbf{u}_{\mathbf{B}}(\theta)^T)^{-1} \quad (16)$$

où $\text{Var}^p(\hat{\mathbf{u}}(\theta))$ est une matrice de dimension $p \times p$ composée des variances des totaux compris dans le vecteur $\hat{\mathbf{u}}(\theta)$ et $\text{Cov}^p(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ est la matrice de variance-covariance des moyennes groupées des covariables paramétriques et de la variable réponse. Les matrices \mathbf{J} et ℓ sont la matrice unitaire de dimension $D \times D$ et le vecteur

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}^{p+1} \otimes \mathbf{A}_1 & 0 & 0 & 0 \\ 0 & \mathbf{I}^{p+1} \otimes \mathbf{A}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{I}^{p+1} \otimes \mathbf{A}_D \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} \text{Cov}^p(\hat{\mathbf{t}}_1, \hat{\mathbf{x}}_1) & \dots & \text{Cov}^p(\hat{\mathbf{t}}_1, \hat{\mathbf{x}}_p) & \text{Cov}^p(\hat{\mathbf{t}}_1, \hat{\mathbf{y}}) \\ \vdots & \ddots & \vdots & \vdots \\ \text{Cov}^p(\hat{\mathbf{t}}_p, \hat{\mathbf{x}}_1) & \dots & \text{Cov}^p(\hat{\mathbf{t}}_p, \hat{\mathbf{x}}_p) & \text{Cov}^p(\hat{\mathbf{t}}_p, \hat{\mathbf{y}}) \end{pmatrix},$$

où, pour $j = 1, \dots, p$, $\hat{\mathbf{t}}_j$ est un vecteur de dimension $D \times 1$ dont la d^e entrée est $\sum_{k \in s_j} w_{jk} u_{jk}(\theta)$ et $\mathbf{A}_d = \mathbf{e}^T(\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_w \mathbf{Z}$ pour $d = 1, \dots, D$.

En remplaçant θ , $\text{Var}^p(\hat{\mathbf{u}}(\theta))$, $\text{Cov}^p(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, \mathbf{A} et \mathbf{C} par leurs estimateurs sur échantillon, nous obtenons l'estimateur par sondage de la variance de $\hat{\mathbf{B}}$:

$$\widehat{\text{Var}}^p(\hat{\mathbf{B}}) = \hat{\mathbf{u}}_{\mathbf{B}}^{-1}(\theta) \\ (\widehat{\text{Var}}^p(\hat{\mathbf{u}}(\theta)) + \mathbf{U}^{\xi}(\theta)(\hat{\mathbf{A}}(\mathbf{I} \otimes \widehat{\text{Cov}}^p(\hat{\mathbf{x}}, \hat{\mathbf{y}}))\hat{\mathbf{A}}^T) \mathbf{U}^{\xi}(\theta)^T \\ + 2(\mathbf{I}^p \otimes \ell)(\ell \otimes \hat{\mathbf{C}})\hat{\mathbf{A}}^T \mathbf{U}^{\xi}(\theta)(\hat{\mathbf{u}}_{\mathbf{B}}(\theta)^T)^{-1},$$

où $\hat{\mathbf{A}}$ est l'estimateur par sondage de \mathbf{A} et est composé de $\hat{\mathbf{A}}_d = \mathbf{e}^T(\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}_w \mathbf{Z}$.

3.3 Propriétés asymptotiques de $\hat{\mathbf{g}}(\cdot)$

Soit $\hat{\mathbf{r}} = \hat{\mathbf{y}} - \hat{\mathbf{x}}\hat{\mathbf{B}}$ l'estimateur sur échantillon de $\mathbf{r} = \mathbf{y} - \mathbf{x}\mathbf{B}$. Une linéarisation autour des paramètres de population, ainsi que l'absence de biais par rapport au plan pour les moyennes de domaine et $\hat{\mathbf{B}}$, aboutissent à l'absence asymptotique de biais par rapport au plan pour $\hat{\mathbf{r}}$. En réexprimant $\hat{\mathbf{g}}(z_d)$, nous obtenons $\hat{\mathbf{g}}(z_d) = \hat{\mathbf{A}}_d \hat{\mathbf{r}}$. Dans $\hat{\mathbf{A}}_d$, nous pouvons développer $(\mathbf{Z}^T \mathbf{K}_w \mathbf{Z})^{-1}$ en utilisant le développement en série de Taylor $(\mathbf{I} + \mathbf{G})^{-1} = \mathbf{I} - \mathbf{G} + \mathbf{G}^2 - \dots$ sachant que \mathbf{G} est une matrice symétrique et inversible. En utilisant les deux premiers termes du développement, nous pouvons montrer que $E^p(\hat{\mathbf{A}}_d)$ est approximativement $\hat{\mathbf{A}}_d$. Donc, nous avons l'absence asymptotique de biais par rapport au plan pour $\hat{\mathbf{g}}(z_d)$. Selon la même méthode, nous obtenons la variance asymptotique approchée par rapport au plan de $\hat{\mathbf{g}}(z_d)$ sous la forme

$$\text{Var}^p(\hat{\mathbf{g}}(z_d)) = \mathbf{A}_d \text{Var}^p(\hat{\mathbf{r}})\mathbf{A}_d^T,$$

où, sachant que $\hat{\mathbf{Q}} = (\mathbf{I}_1 - \mathbf{B}_1, \dots, -\mathbf{B}_p)$,

rapport à $\mathbf{m}_\xi(\mathbf{z})$. Désignons par $\mathbf{u}_B(\theta)$ et $\mathbf{U}_\xi(\theta)$ les

paramètres de population correspondant à $\mathbf{u}_B(\theta)$ et $\mathbf{U}_\xi(\theta)$, respectivement.

En plus des conditions de régularité susmentionnées,

nous imposons les conditions suivantes, en désignant par

\mathcal{N} un voisinage de la valeur réelle des paramètres d'intérêt.

C1. $\lim_{N \rightarrow \infty} \mathbf{u}(\theta)/N$ existe et est finie pour tous θ et

C2. $\lim_{N \rightarrow \infty} \mathbf{u}_B(\theta)/N = \mathbf{H}_B$ et \mathbf{H}_B est de plein rang

et inversible pour tous θ et \mathcal{N} .

C3. $\lim_{N \rightarrow \infty} \mathbf{U}_\xi(\theta)/N = \mathbf{H}_\xi(\theta)$ et $\mathbf{H}_\xi(\theta)$ possède un

déterminant fini pour tous θ et \mathcal{N} .

C4. $\lim_{N \rightarrow \infty} n \text{Var}^p(\mathbf{u}(\theta)/N) = \mathbf{V}(\mathbf{u}(\theta))$, où Var^p

est une matrice de variance définie positive pour

tous θ et \mathcal{N} .

C5. $\lim_{N \rightarrow \infty} N^p/N = \omega_p$ et $\lim_{N \rightarrow \infty} n/N = f$, où

ω_p ainsi que f sont des constantes comprises

entre 0 et 1.

C6. Soit $\mathbf{A}_d = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w$ la matrice de

lissage de population; alors $\lim_{N \rightarrow \infty} \mathbf{A}_d$ existe et

est finie pour $d = 1, \dots, D$.

C7. $\lim_{N \rightarrow \infty} n \text{Var}^p(\mathbf{m}_\xi(\mathbf{z})) = \mathbf{V}(\mathbf{m}_\xi(\mathbf{z}))$.

C8. Les matrices des valeurs de population $\mathbf{Z}^T \mathbf{K}^w \mathbf{Z}$

et $\mathbf{u}_B(\theta)$ sont inversibles, ainsi que leurs estima-

teurs par échantillonnage $\mathbf{Z}^T \mathbf{K}^w \mathbf{Z}$ et $\mathbf{u}_B(\theta)$.

3.2 Propriétés asymptotiques de B

Les preuves de tous les lemmes et théorèmes exposés

dans la présente section et dans la suivante figurent en

annexe. D'après les résultats de la linéarisation de Taylor

dans (13), nous savons que les propriétés de \mathbf{B} dépendent

de celles de $\mathbf{u}(\theta)$, $\mathbf{u}_B(\theta)$, $\mathbf{U}_\xi(\theta)$ et $\mathbf{m}_\xi(\mathbf{z})$; leurs pro-

priétés sont énoncées dans les deux lemmes qui suivent.

Lemme 1. Si les conditions C1 à C4 sont satisfaites,

nous avons, quand $v \rightarrow \infty$:

$$(1) \sqrt{n}(\mathbf{u}(\theta)/N - \mathbf{u}(\theta))/N \longrightarrow N(0, \mathbf{V}(\mathbf{u}(\theta))) ;$$

$$(2) \|\mathbf{u}_B(\theta) - \mathbf{u}_B(\theta)\|/N \text{ et } \|\mathbf{U}_\xi(\theta) - \mathbf{U}_\xi(\theta)\| \text{ convergent vers } 0 \text{ en probabilité pour } \theta \text{ et } \mathcal{N} ;$$

$$(3) \|\mathbf{u}(\theta)/N - \mathbf{u}(\theta)\|/N \text{ converge vers zéro en probabilité.}$$

Lemme 2. Sous les conditions C5 à C7, $\sqrt{n}(\mathbf{m}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z}))/O^p(1)$.

En nous appuyant sur les lemmes 1 et 2, nous obtenons la normalité asymptotique de \mathbf{B} dans le théorème 1.

Théorème 1. Sous les conditions C1 à C7, en supposant que l'espace des paramètres contient un voisinage du paramètre d'intérêt, nous avons, quand v tend vers l'infini :

$$\mathbf{M}_y = \begin{pmatrix} m_y(z_1) & \vdots & m_y(z_1) \\ m_y(z_1) & \vdots & m_y(z_d) \\ m_y(z_d) & \vdots & m_y(z_d) \end{pmatrix}$$

Soit n_d le nombre d'observations dans la d^e classe, tel que $\sum n_d = n$. Similairement à \mathbf{M}_x et \mathbf{M}_y dans (7), les d^e blocs de \mathbf{M}_x et \mathbf{M}_y sont de dimensions $n_d \times p$ et $n_d \times 1$, respectivement.

Par analogie avec l'équation d'estimation en population (8), l'équation d'estimation sur échantillon pour \mathbf{B} est

$$\hat{\mathbf{u}}(\hat{\theta}) = \sum_{k \in s} (\mathbf{x}_k - \hat{\mathbf{M}}_x^k)^T (\mathbf{y}_k - \hat{\mathbf{M}}_y^k) \mathbf{w}_k - \sum_{k \in s} (\mathbf{x}_k - \hat{\mathbf{M}}_x^k)^T (\mathbf{x}_k - \hat{\mathbf{M}}_x^k) \mathbf{B} \mathbf{w}_k = \mathbf{0}, \quad (12)$$

où $\hat{\theta}^T = (\hat{\mathbf{B}}^T, \hat{\mathbf{m}}_x(z), \hat{\mathbf{m}}_y(z)^T)^T$ est l'estimateur par échantillonnage de $\theta^T = (\mathbf{B}^T, \mathbf{m}_x(z), \mathbf{m}_y(z)^T)^T$. Notons qu'une approche semblable a été envisagée par Fuller (1975) et par Binder (1983). Néanmoins, la solution de (12) donne comme forme explicite de \mathbf{B} l'expression

$$\hat{\mathbf{B}} = ((\mathbf{x} - \hat{\mathbf{M}}_x)^T \mathbf{w} (\mathbf{x} - \hat{\mathbf{M}}_x)^{-1} (\mathbf{x} - \hat{\mathbf{M}}_x)^T \mathbf{w} (\mathbf{y} - \hat{\mathbf{M}}_y)),$$

où \mathbf{W}_n est une matrice de poids de dimensions $n \times n$ avec les poids de sondage w_k sur la diagonale pour $k \in s$, \mathbf{y} est un vecteur de dimension $n \times 1$ contenant les observations sur échantillon de la variable réponse et \mathbf{x} est une matrice de dimensions $n \times p$ contenant les observations sur échantillon des covariables.

En nous servant des estimations sur échantillon de \mathbf{B} et en désignant par $\hat{\mathbf{x}}$ une matrice de dimensions $D \times p$ de la forme $(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_p)$, nous pouvons obtenir l'estimation sur échantillon de $g(z_d)$ sous la forme

$$\hat{g}(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w (\mathbf{Y} - \hat{\mathbf{x}} \mathbf{B}).$$

De nouveau, si g et h sont les mêmes que pour $m_j(z_d)$, l'expression donnant $\hat{g}(z_d)$ se simplifie.

Quand nous appliquons les méthodes de régression polynomiale locale pour obtenir les estimateurs des espérances conditionnelles ainsi que la fonction arbitraire $G(\cdot)$, nous devons choisir une largeur de fenêtre h appropriée. Comme le groupement par classe intervient dans tous les aspects du processus d'estimation et que nous supposons que les classes induites par les valeurs distinctes de \mathbf{z} sont

préservées lorsqu'on passe de la population à l'échantillon, nous soutenons que la même largeur de fenêtre devrait être utilisée pour obtenir les estimateurs sous recensement et les estimateurs sous échantillonnage. Puisque nous ne disposons pas de toutes les observations de la population finie, nous utilisons l'échantillon pour choisir la largeur de fenêtre appropriée. Dans le présent article, nous adoptons la méthode de Fan et Gijbels (1995), qui ont élaboré un sélecteur de largeur de fenêtre dicté par les données qui combine les idées des méthodes d'insertion de données et de validation croisée pour les données indépendantes et identiquement distribuées. Afin d'appliquer cette méthode dictée par les données à notre cas, nous avons besoin de critères, tels que la somme des carrés des résidus et l'erreur quadratique moyenne, concernant les estimations résultantes des espérances conditionnelles. En notant que ces critères dépendent des espérances conditionnelles estimées ou des fonctions de régression et des dérivées des fonctions de régression, nous pouvons utiliser les fonctions objectif définies en (9) et (10) pour obtenir non seulement les estimations par sondage des fonctions de régression, mais aussi les dérivées de ces fonctions. Pour plus de détails, voir Wang (2004).

3. Propriétés sous le plan des estimateurs par échantillonnage

3.1 Notation et hypothèses

Afin d'illustrer les propriétés des estimateurs sous le plan d'échantillonnage, à l'instar de Särndal, Swensson et Wretman (1992) et d'Isaki et Fuller (1982), nous considérons une série de populations emboîtées U_v , pour $v = 1, 2, \dots$, telles que $U_1 \subset U_2 \subset \dots \subset U_v$. Toutes les grandeurs de population, les tailles et valeurs d'échantillon et les estimateurs par sondage ont l'indice v . Cependant, pour simplifier la notation, nous laissons tomber l'indice inférieur v pour ces quantités. Nous désignons l'espérance et la variance par rapport au plan d'échantillonnage par E_p et Var_p , respectivement, et conformément aux populations emboîtées susmentionnées, nous définissons la convergence par rapport au plan et l'absence asymptotique de biais comme dans Thompson (1997, page 167).

Dans la suite de l'exposé, le développement des résultats asymptotiques pour les estimateurs dépendra de la normalité et de la convergence des estimations des espérances moyennes et des totaux. Nous ne nous limiterons pas à des plans d'échantillonnage particuliers; au contraire, nous supposons que tous les totaux d'enquête qui figurent dans les estimateurs sont de type Horvitz-Thompson. Donc, la convergence et la normalité asymptotique des estimateurs sont soumises aux conditions de régularité standard appliquées aux plans d'échantillonnage pour la convergence et

obtenir les versions en population finie des paramètres (estimateurs sous recensement) de β , nommément \mathbf{B} , en résolvant

$$\mathbf{u}(\theta) = \sum_{k=1}^K (\mathbf{x}_k - \mathbf{M}^{\mathbf{x}})^T (\mathbf{y}_k - M^{\mathbf{y}}) - \sum_{k=1}^K (\mathbf{x}_k - \mathbf{M}^{\mathbf{x}})^T (\mathbf{x}_k - \mathbf{M}^{\mathbf{x}}) \mathbf{B} = \mathbf{0}^{p \times 1}$$

(8)

où $\mathbf{M}^{\mathbf{x}}$ est la k^e ligne de la matrice $\mathbf{M}^{\mathbf{x}}$ de dimensions $N \times p$ et $M^{\mathbf{y}}$ est le k^e élément du vecteur $\mathbf{M}^{\mathbf{y}}$ de dimension $N \times 1$. Le vecteur des paramètres de population finie θ^* est composé de $(\mathbf{B}^T, \mathbf{m}_1(z), \mathbf{m}_y(z)^T, \text{ou } \mathbf{m}_1(z)$ avec $\mathbf{m}_j(z) = (m_j(z_1), \dots, m_j(z_D))$ pour $j = 1, \dots, p$ et $\mathbf{m}_y(z) = (m_y(z_1), \dots, m_y(z_D))$. D'où l'expression explicite pour l'estimateur (paramètre sous recensement) \mathbf{B} est

$$\mathbf{B} = ((\mathbf{X} - \mathbf{M}^{\mathbf{x}})^T (\mathbf{X} - \mathbf{M}^{\mathbf{x}}))^{-1} (\mathbf{X} - \mathbf{M}^{\mathbf{x}})^T (\mathbf{Y} - \mathbf{M}^{\mathbf{y}}).$$

Une fois que \mathbf{B} est obtenu, la différence entre la variable réponse \mathbf{Y} et le produit $\mathbf{X}\mathbf{B}$ est traitée comme la variable aléatoire dépendante et la fonction $G(\cdot)$ est estimée conformément au modèle suivant

$$\mathbf{Y} - \mathbf{X}\mathbf{B} = G(\mathbf{z}) + \mathbf{e}.$$

La version en population finie de $G(\mathbf{z})$ à z_d , nommément

$$g(z_d), \text{ est}$$

$$g(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w (\mathbf{Y} - \mathbf{X}\mathbf{B}),$$

où \mathbf{X} est une matrice de dimensions $D \times p$ de la forme

$$(\mathbf{X}_1, \dots, \mathbf{X}_p),$$

En réalité, nous n'avons pas accès à l'ensemble de la population. Au contraire, nous ne pouvons observer qu'un échantillon tiré de la population en utilisant un certain plan d'échantillonnage probabiliste. Soit s l'ensemble de n

unités échantillonnées avec l'échantillon $(y_k, \mathbf{x}_k, z_k, w_k)$ pour $k \in s$, où w_k est le poids d'échantillonnage de l'unité k . En outre, nous supposons que la réponse est complète, si bien que la probabilité d'inclusion est égale à l'inverse du poids d'échantillonnage. Nous supposons aussi que les classes induites par les valeurs distinctes de \mathbf{z} sont préservées lorsque l'on passe de la population à l'échantillon. Cette hypothèse est appropriée dans le cas d'une variable telle que l'âge enregistré comme étant l'âge au dernier anniversaire.

En appliquant la méthode de régression polynomiale locale pour données d'enquêtes complexes de Bellhouse et

$$\sum_{d=1}^p \frac{h}{\bar{p}_d} \{ \underline{\mathbf{y}}^T \bar{p}_d^p - \alpha_0 - \alpha_1 (z_d^p) - \dots - \alpha_b (z_d^p) \} \times \left(K \frac{h}{z_d^p - z_d^p} \right) \quad \text{et} \quad \sum_{d=1}^p \frac{h}{\bar{p}_d} \{ \underline{\mathbf{x}}^T \bar{p}_d^p - \gamma_0 - \gamma_1 (z_d^p) - \dots - \gamma_b (z_d^p) \} \times \left(K \frac{h}{z_d^p - z_d^p} \right), \quad (10)$$

où $\underline{\mathbf{y}}$ et $\underline{\mathbf{x}}$, sont des estimateurs sur échantillon de \mathbf{Y} et \mathbf{X} et sont de la forme $(\underline{y}_1, \dots, \underline{y}_D)^T$ et $(\underline{x}_1, \dots, \underline{x}_D)^T$, respectivement, et \bar{p}_d est la proportion pondérée, dans l'échantillon, des observations comprises dans la classe d . Par conséquent, les estimateurs par sondage de $m_y(z)$ et $m_f(z)$ à z_d sont donnés par

$$\hat{m}_f(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w \underline{\mathbf{x}}$$

et

$$\hat{m}_y(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w \underline{\mathbf{y}},$$

où \mathbf{Z} est de la même forme que dans (6) et \mathbf{K}^w est défini comme étant

$$\mathbf{K}^w = \frac{1}{I} \text{diag}(\hat{p}_1 K((z_1 - z_d)/h), \dots, \hat{p}_D K((z_D - z_d)/h)).$$

Nous pouvons aussi construire la matrice $\mathbf{M}^{\mathbf{x}}$ de dimensions $n \times p$ et le vecteur $\mathbf{M}^{\mathbf{y}}$ de dimension $n \times 1$ par la même méthode que celle utilisée pour construire $\mathbf{M}^{\mathbf{x}}$ et $\mathbf{M}^{\mathbf{y}}$ dans les équations (7). Autrement dit, nous utilisons les estimateurs par échantillonnage $\hat{m}_f(z_d)$ et $\hat{m}_y(z_d)$ donnés en (11) pour obtenir

$$\mathbf{M}^{\mathbf{x}} = \begin{pmatrix} m_{x_1}(z_1) & m_{x_2}(z_1) & \dots & m_{x_p}(z_1) \\ m_{x_1}(z_2) & m_{x_2}(z_2) & \dots & m_{x_p}(z_2) \\ \vdots & \vdots & \ddots & \vdots \\ m_{x_1}(z_D) & m_{x_2}(z_D) & \dots & m_{x_p}(z_D) \end{pmatrix}$$

distinctes. En outre, nous imaginons que $E(\mathbf{e} | \mathbf{z}, \mathbf{X}) = \mathbf{0}$. Il n'existe aucune interaction entre \mathbf{X} et \mathbf{z} dans le modèle. Nous voulons estimer les versions au niveau de la population de $G(\cdot)$ et des paramètres β . Nous commençons par élaborer des expressions pour ces entités, en nous inspirant des méthodes d'estimation décrites dans Robinson (1988) et dans Speckman (1988). En particulier, nous commençons par prendre l'espérance des deux membres de (1) sachant \mathbf{z} :

$$(2) \quad E(\mathbf{Y} | \mathbf{z}) = E(\mathbf{X} | \mathbf{z})\beta + G(\mathbf{z}).$$

Puis, nous soustrayons (2) de (1) pour obtenir

$$(3) \quad \mathbf{Y} - E(\mathbf{Y} | \mathbf{z}) = (\mathbf{X} - E(\mathbf{X} | \mathbf{z}))\beta + \mathbf{e}.$$

Pour définir la version de population de β dans (3), nous remplaçons $E(\mathbf{Y} | \mathbf{z})$ et $E(\mathbf{X} | \mathbf{z})$ dans (3) par leurs estimations au niveau de la population et nous estimons β par la méthode des moindres carrés.

Pour les estimations au niveau de la population de $E(\mathbf{Y} | \mathbf{z})$ et $E(\mathbf{X} | \mathbf{z})$, nous adoptons le lissage à polynômes locaux de Jones (1989), dans lequel le groupement

par classe est un élément essentiel de l'opération. Soit la variable discrétisée Z qui prend les valeurs z_1, \dots, z_D ; soit $\mathbf{Y} = (\mathbf{Y}^1, \dots, \mathbf{Y}^D)$ et $\mathbf{X} = (\mathbf{X}^1, \dots, \mathbf{X}^D)$ pour $j = 1, \dots, D$. Alors, désignons les espérances conditionnelles lissées de population de \mathbf{Y} et \mathbf{X}_j au point z_d par $m_j(z_d)$ et $m_j(z_d)$, respectivement. Sachant que $K(\cdot)$ est une fonction moyen qui satisfait $\int K(t) dt = 1$ et $\int K(t) dt < \infty$ et que h est la largeur de fenêtre, et en utilisant le principe de la méthode de régression polynomiale locale, nous minimisons

$$\sum_{d=1}^D \frac{h}{p} \{ \mathbf{Y}^d - \alpha_0 - \alpha_1(z_d' - z_d), \dots, -\alpha_b(z_d' - z_d) \}^2$$

$$(4) \quad \times K \left(\frac{h}{z_d' - z_d} \right)$$

et

$$\sum_{d=1}^D \frac{h}{p} \{ \mathbf{X}^d - \gamma_0 - \gamma_1(z_d' - z_d), \dots, -\gamma_b(z_d' - z_d) \}^2$$

$$(5) \quad \times K \left(\frac{h}{z_d' - z_d} \right)$$

par rapport aux α et aux γ de sorte que les espérances conditionnelles lissées (lissées) de population de \mathbf{Y} et \mathbf{X}_j sur z_d , $m_j(z_d)$ et $m_j(z_d)$ soient les solutions de α_0 et γ_0 pour les équations (4) et (5). En particulier,

$$m_y(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w \mathbf{Y}$$

et

$$m_j(z_d) = \mathbf{e}^T (\mathbf{Z}^T \mathbf{K}^w \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{K}^w \mathbf{X}_j$$

où q est le degré du lisseur polynomial, \mathbf{e} est un vecteur de dimension $(q+1) \times 1$ de la forme $(1, 0, 0, \dots, 0)^T$, et \mathbf{Z} et \mathbf{K}^w sont définis respectivement par

$$(6) \quad \mathbf{Z} = \begin{pmatrix} 1 & z_1 - z_d & \dots & (z_1 - z_d)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_2 - z_d & \dots & (z_2 - z_d)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_D - z_d & \dots & (z_D - z_d)^q \end{pmatrix}$$

et $\mathbf{K}^w = \text{diag}(P_1 K((z_1 - z_d)/h), \dots, P_D K((z_D - z_d)/h))$. Au moyen des estimateurs sous recensement des espérances conditionnelles $m_j(z_d)$ et $m_j(z_d)$, nous définissons une matrice \mathbf{M}_x de dimensions $N \times p$ et un vecteur \mathbf{M}_y de dimension $N \times 1$ sous la forme

$$\mathbf{M}_x = \begin{pmatrix} m_1(z_1) & m_2(z_1) & \dots & m^d(z_1) \\ \vdots & \vdots & \ddots & \vdots \\ m_1(z_D) & m_2(z_D) & \dots & m^d(z_D) \end{pmatrix}$$

(7)

$$\mathbf{M}_y = \begin{pmatrix} m_y(z_1) \\ \vdots \\ m_y(z_D) \end{pmatrix}$$

Notons que les d^e blocs de \mathbf{M}_x et \mathbf{M}_y sont de dimensions $N_d \times p$ et $N_d \times 1$, respectivement, où N_d est le nombre d'observations qui rentrent dans d^e classe et $\sum N_d = N$. En remplaçant la matrice, $E(\mathbf{X} | \mathbf{z})$, et le vecteur, $E(\mathbf{Y} | \mathbf{z})$, des espérances conditionnelles dans (3) par leurs estimations, \mathbf{M}_x et \mathbf{M}_y , et en utilisant le cadre des équations d'estimations généralisées proposées par Godambe et Thompson (1986) pour l'estimation des moindres carrés, nous pouvons

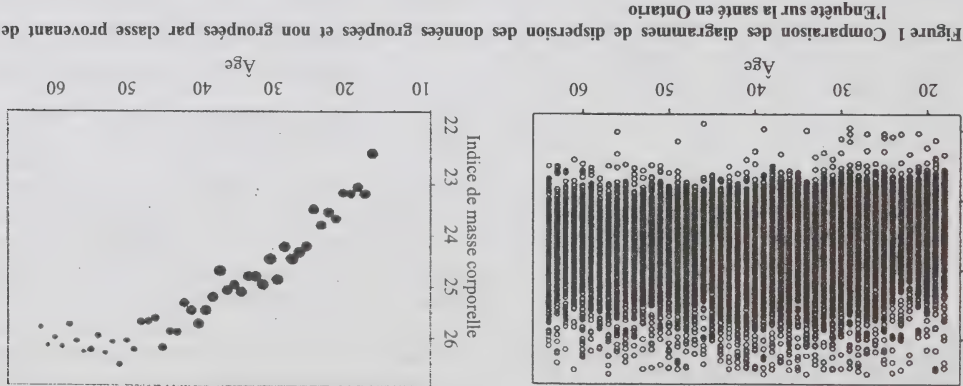


Figure 1 Comparaison des diagrammes de dispersion des données groupées et non groupées par classe provenant de l'Enquête sur la santé en Ontario

2. Modèle de régression semiparamétrique et son estimation

Nous adoptons une approche typique de l'analyse des données d'enquêtes complexes. Pour commencer, nous considérons un modèle de travail appliqué à la population finie sous l'hypothèse que les observations sont indépendantes. Les estimations des paramètres du modèle de- viennent alors les paramètres de population finie, ou para- mètres sous recensement, qui doivent être estimés d'après l'échantillon. Une fois que nous avons défini les paramètres cibles de population finie, nous considérons un modèle hypothétique plus réaliste de la population finie afin d'ob- tenir des inférences au sujet de ces paramètres, ce que nous faisons à la section suivante. Considérons une population finie de taille N avec un vecteur de mesures (y_k, x_k, z_k) attaché à l'unité k , $k = 1, \dots, N$, où y_k représente une observation de la variable réponse et (x_k, z_k) représente un vecteur d'observations des variables explicatives de dimen- sion $p + 1$. À titre de modèle de travail, nous imaginons que la variable réponse est produite par le modèle de ré- gression linéaire partiel suivant

$$Y = G(Z) + X\beta + \varepsilon \quad (1)$$

où Y est le vecteur de réponses et ε contient des entrées indépendantes et identiquement distribuées de moyenne nulle et de variance constante. La fonction $G(\cdot)$ est une fonction arbitraire de z et β est un vecteur de paramètres de dimension p inconnu. La matrice X de dimensions $N \times p$ correspond à la partie linéaire du modèle et contient des variables explicatives continues ou discrètes qui sont aléatoires. Le terme $G(Z)$ est la partie non paramétrique et modèle. Nous supposons que z est non stochastique et mesurée sur une échelle continue, discrétisée en D valeurs

Un inconvénient du groupement par classe est que le nombre de classes ne peut pas croître asymptotiquement avec la population si les données sont groupées naturel- lement, comme cela est le cas de la variable d'âge dans l'exemple susmentionné. Dans de telles conditions, les estimateurs non paramétriques au niveau de la population demeurent entachés d'un biais en tant qu'estimateurs de fonctions de superpopulation, en raison de la taille de classe fixe. Dans notre cadre, nous supposons que les classes induites par les valeurs distinctes de la covariable sont les mêmes dans la population que dans l'échantillon ; de même, dans le lissage, nous supposons que la largeur de fenêtre est la même au niveau de la population qu'au niveau de l'échantillon. Nous montrerons que les estimateurs sur l'échantillon sont des estimateurs convergents par rapport au plan des paramètres et fonctions de population finie cor- respondants, mais non de leurs analogues en superpopu- lation. Dans l'exemple des données de l'Enquête sur la santé en Ontario, le même ensemble d'âges distincts s'observe dans la population et dans l'échantillon.

La présentation de l'article est la suivante. À la section 2, nous introduisons les modèles de travail en superpopulation qui mènent aux méthodes d'estimation dans le cas de données d'enquête. À la section 3, nous calculons tous les moments des estimations obtenues et établissons certains résultats asymptotiques. Aux sections 4 et 5, nous présen- tons une étude par simulation et un exemple empirique de la méthode d'estimation appliquée en utilisant les données de l'Enquête sur la santé en Ontario de 1990 (1992). À la section 6, nous concluons par une discussion des hypothèses formulées et de certains futurs travaux. Les preuves de tous les lemmes et théorèmes présentés à la section 3 sont données en annexe.

Un exemple qui illustre ces caractéristiques des données groupées par classe est tiré de l'Enquête sur la santé en Ontario. Cette enquête a été réalisée par Statistique Canada en 1990 auprès de 61 239 personnes vivant en Ontario, au Canada. Les données ont été obtenues au moyen d'un plan d'échantillonnage en grappes stratifié à deux degrés. Les strates correspondaient aux régions urbaines et rurales relevant de chacun des bureaux de santé de l'Ontario. Des secteurs de dénombrement ont été sélectionnés aléatoirement dans chaque strate et, de même, des ménages l'ont été dans chaque secteur de dénombrement. L'objectif de cette enquête est de mesurer l'état de santé des habitants de l'Ontario et de recueillir des données sur les facteurs de risque associés aux principales causes de mortalité et de morbidité dans la province. Dans notre exemple, nous examinons le poids de la personne en fonction de l'âge. Dans l'Enquête sur la santé en Ontario, l'âge déclaré est celui au dernier anniversaire. La mesure que nous utilisons ici comme substitut du poids est l'indice de masse corporelle (IMC), qui est calculé en divisant le poids exprimé en kilogrammes par le carré de la taille exprimée en mètres. L'IMC est l'un des indicateurs du degré d'obésité considéré comme une insuffisance pondérale et un IMC supérieur à 30, comme une indication d'obésité. L'IMC n'est une mesure appropriée que pour les personnes de 18 à 64 ans, à l'exception des femmes enceintes ou qui allaient. Par conséquent, la taille de l'échantillon est réduite à 44 457 répondants admissibles qui peuvent être répartis entre 47 âges ou classes.

Le graphique de gauche de la figure 1 représente la tendance de l'indice de masse corporelle en fonction de l'âge. Il est facile de voir que le diagramme de dispersion semblable à un « nuage noir » masque la relation entre l'âge et l'indice de masse corporelle. Par contre, si nous calculons la moyenne de l'indice de masse corporelle à chaque point d'âge distinct et que nous représentons graphiquement les estimations moyennes par classe de l'indice de masse corporelle en fonction de l'âge, nous obtenons le graphique de droite de la figure 1. Il est évident qu'une moyenne groupée par classe fournit plus de renseignements visuels que les données brutes. Les grands ensembles de données peuvent non seulement donner lieu à des graphiques très informatifs, mais aussi rendre le calcul des estimations très fastidieux. Donc, il est naturel, dans l'analyse des données complexes, de regrouper les données dans des domaines en fonction des valeurs distinctes d'une covariable discrétisée. En outre, les estimateurs résultant du groupement par classe sont des fonctions des estimateurs de domaine.

indépendantes paramétriques, sachant la variable non paramétrique, sont traitées comme une fonction de cette variable et lissées : à la deuxième étape, les coefficients linéaires sont estimés par régression des résidus provenant de la variable réponse lissée sur les résidus provenant des covariables paramétriques lissées. Enfin, la différence entre la variable réponse et sa prédiction d'après le modèle de régression est lissée de façon semblable pour produire une estimation de la partie non paramétrique de la fonction de régression. Robinson (1988) et Speckman (1988) ont montré que les estimateurs résultants sont convergents à la racine carrée de n quand le modèle est correct et que les points de données sont indépendants et identiquement distribués. L'objectif de notre article est d'appliquer cette méthode de lissage à des données d'enquête tout en tenant compte d'un plan d'échantillonnage complexe.

Nous utilisons la méthode d'estimation par la régression polynomiale locale établie dans Bellhouse et Stafford (2001) pour effectuer le lissage durant le processus d'estimation. Un élément clé de l'exécution de cette méthode est le groupement par classe ou fenêtre (*binning*), qui découle des travaux de Bellhouse et Stafford (1999) sur l'estimation de la densité. Dans de nombreux ensembles de données d'enquête, une variable continue peut être naturellement groupée par classe ; par exemple, l'âge peut être enregistré comme l'âge au dernier anniversaire. En général, les classes ou fenêtres correspondent aux ensembles disjoints de valeurs d'une covariable continue et, par conséquent, peuvent être considérées comme des domaines. Au niveau de l'échantillon, nous estimons la moyenne de domaine de la variable d'intérêt en divisant la somme pondérée de la variable dans le domaine par la somme des poids dans le domaine. Dans Bellhouse et Stafford (2001), la variable réponse est groupée par classe en fonction des valeurs de la covariable et les moyennes de domaine de la variable réponse sont lissées pour obtenir la fonction de régression. Quand la taille d'échantillon est grande et que le nombre de classes est relativement faible, les estimateurs basés sur le groupement par classe sont des fonctions des estimateurs de domaine dont les propriétés inférencielles peuvent être facilement établies d'après les résultats présentés dans Shao (1996) et dans Serfling (1980). L'un des avantages pratiques du groupement par classe est qu'il peut révéler l'information sur une tendance qui est obscurcie dans une enquête complexe et qui peut être importante si l'ensemble de données d'enquêtes complexes est très grand. Habituellement, il existe de multiples observations pour chaque ensemble de valeurs des covariables dans ces ensembles de données.

Modèle de régression semiparamétrique pour les données d'enquêtes complexes

Zilin Wang et David R. Bellhouse¹

Résumé

Nous élaborons un modèle de régression semiparamétrique pour les enquêtes complexes. Dans ce modèle, les variables explicatives sont représentées séparément sous forme d'une partie non paramétrique et d'une partie linéaire paramétrique. Les méthodes d'estimation combinent l'estimation par la régression polynomiale locale non paramétrique et l'estimation par les moindres carrés. Nous élaborons également des résultats asymptotiques, tels que la convergence et la normalité des estimateurs des coefficients de régression et des fonctions de régression. Nous recourons à la simulation et à des exemples empiriques tirés de l'Enquête sur la santé en Ontario de 1990 pour illustrer la performance de la méthode et les propriétés des estimations.

Mots clés : Enquête complexe ; estimation par domaine ; régression non paramétrique ; lissage.

1. Introduction

En pratique, nombre d'enquêtes sont utilisées pour étudier la relation entre une variable réponse et des variables explicatives, ainsi que pour constituer des modèles prédictifs. Par conséquent, il est nécessaire de mettre au point des méthodes qui permettent d'appliquer les modèles de régression stochastiques à des données d'enquête. Alors que les méthodes de régression non paramétrique sont utilisées largement dans de nombreux domaines de la statistique, peu d'attention leur a été accordée dans celui des enquêtes complexes, à cause de la complexité de la structure des données. Étant donné la corrélation due au tirage de l'échantillon avec mise en grappes et les probabilités de sélection inégales, les données de ces enquêtes ne sont ni indépendantes ni identiquement distribuées. Les méthodes de régression non paramétrique standard sont donc souvent inappropriées pour l'analyse des données d'enquête par sondage.

Certains auteurs, par exemple Breidt et Opsomer (2000),

Montanari et Ranalli (2005), et Zheng et Little (2004), se sont penchés sur l'élaboration de méthodes de régression non paramétrique applicables aux données d'enquête. Cependant, comme dans le cas de l'application classique des méthodes de régression, la plupart de ces travaux s'appuient sur des approches assistées par modèle pour estimer les grands paramètres de population descriptives et les paramètres reliés à ces grands paramètres. Dans le présent article, nous nous intéressons à l'application de méthodes de régression non paramétrique pour étudier la relation entre la variable réponse et les covariables, ainsi que la prédiction en utilisant l'information auxiliaire. Bellhouse et Stifford (2001) ont étendu une méthode de régression polynomiale

linéaire dans la partie paramétrique du modèle, $G(\mathbf{z})$, et une variable fournissant peu d'information sur la forme fonctionnelle dans la partie non paramétrique, $G(\mathbf{z})$. Non seulement ce modèle est dicté par la motivation a priori d'en faire un outil d'analyse des données et retient une fonction d'interprétation importante, mais il facilite aussi la résolution du problème de dimensionnalité élevée créé par les facteurs et certaines covariables, grâce à leur inclusion dans la partie paramétrique du modèle.

Un modèle semblable a été élaboré pour des données indépendantes et identiquement distribuées par Robinson (1988) et par Speckman (1988). Dans ces articles, l'estimation est effectuée en trois étapes. À la première étape, les moyennes de la variable réponse et des variables

¹ Zilin Wang, Département de mathématiques, Université Wilfrid Laurier, Waterloo (Ontario) Canada, N2L 3C5. Courriel : zwang@wlu.ca; David R. Bellhouse, Département de sciences statistiques et actuariales, Université Western Ontario, Londres (Ontario) Canada, N6A 5B7. Courriel : bellhouse@stats.uwo.ca

ou $\text{Var}(L^{-1}d_{WS}) = a - b + c$. Si L est grand, $\text{Var}(f_{L,ms}^{L,ms}) = (m_0 - 1)a - (m_0 - 1)b + \{3(m_0 + 2)\}^{-1}4(m_0 - 1)^2$. Donc, pour L grand, $V(f_{L,ms}^{L,ms}) = V(f_{L,ms}^{L,ms}) = (m_0 - 2)b + \{3(m_0 + 2)\}^{-1}(4m_0 - 1)m_0 - 2$. Par conséquent, $\lim_{L \rightarrow \infty} V_{L \rightarrow \infty}^{L \rightarrow \infty} \{1c + \sqrt{144a^2 + 144b^2 + 153c^2 - 288ab + 216ac - 216bc}\}$. En particulier, la différence $\lim_{L \rightarrow \infty} V_{L \rightarrow \infty}^{L \rightarrow \infty} - \lim_{L \rightarrow \infty} V_{L \rightarrow \infty}^{L \rightarrow \infty}$ devient égale ou supérieure à zéro quand $m_0 = 10$, quelles que soient les valeurs de a , b et c . Parce qu'il s'agit d'une fraction croissante en m_0 , pour toutes valeurs de $m_0 \geq 10$, $\lim_{L \rightarrow \infty} V_{L \rightarrow \infty}^{L \rightarrow \infty} \geq \lim_{L \rightarrow \infty} V_{L \rightarrow \infty}^{L \rightarrow \infty}$.

Bibliographie

Casady, R., Dorfman, A.H. et Wang, S. (1998). Intervalles de confiance des paramètres de domaine et la taille de l'échantillon du domaine est aléatoire. *Techniques d'enquête*, 24, 59-69.

Cochran, G.C. (1977). *Sampling Techniques*, 3^{ème} Ed. New York : John Wiley & Sons, Inc.

Ellingre, J.L., et Jang, D. (1996). Mesures de la stabilité des échantillons stratifiés à plusieurs degrés. *Techniques d'enquête*, 22, 159-168.

Fuller, W.A. (1987). *Measurement Error Models*. New York : John Wiley & Sons, Inc.

Isaki, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78, 117-123.

Jang, D. (1996). *Stability of Variance Estimators Under Complex Sampling Designs*. Thèse de doctorat non-publiée, Department of Statistics, Texas A&M University, College Station, Texas.

Kendall, M., Stuart, A. et Ord, J.K. (1983). *The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time Series*. New York : Macmillan.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.*, 9, 1010-1019.

National Center for Health Statistics (1996). NHANES III Reference Manuals and Reports, CD-ROM GPO, 017-022-1358-4. Washington, D.C. : United States Government Printing Office.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

Smith, H.F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, 9, 211-212.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.

Les principaux résultats présentés aux sections 1 à 3 peuvent facilement être étendus à partir des variances à l'intérieur des unités primaires d'échantillonnage V_{wh}^h à des variables auxiliaires plus générale X_h . Dans ce genre d'extension, les principaux problèmes demeurent l'adéquation de l'approximation de proportionnalité (1.6) et l'importance de l'erreur d'échantillonnage dans les estimateurs auxiliaires X_h , disons, relativement à l'erreur dans l'estimateur élémentaire de variance au niveau de la strate V_h .

Remerciements

Les auteurs remercient le National Center for Health Statistics des États-Unis de leur avoir donné accès à l'ensemble de données de la NHANES III et remercient V.L. Parsons, C. Johnson et L.R. Curtin d'avoir partagé avec eux une foule de renseignements concernant la NHANES III. La présente étude a été financée en partie par le National Center for Health Statistics des États-Unis. Les opinions exprimées dans le présent article sont celles des auteurs et ne représentent pas forcément les politiques du National Center for Health Statistics des États-Unis ni celles du Bureau of Labor Statistics des États-Unis.

Annexe A

Preuve du résultat 2.2

Considérons une fonction non linéaire $B^{-1}A^2$ de deux estimateurs A et B de moyennes μ_A et μ_B , respectivement. Alors, la variance du premier terme d'un développement en série de Taylor de $B^{-1}A^2$ est

$$\frac{4\mu_A^2}{2}\text{Var}(A) - 4\frac{\mu_A^3}{3}\text{Cov}(A, B) + \frac{\mu_A^4}{4}\text{Var}(B). \quad (\text{A.1})$$

Maintenant, définissons les deux estimateurs suivants : $L^{-1}d_{s1} = B^{-1}A_1^2$ et $L^{-1}d_{s2} = B^{-1}A_2^2$, où $A_1 = L^{-1}\sum_{h=1}^H V_h$ et $B_1 = L^{-1}\sum_{h=1}^H V_h^2$, $A_2 = L^{-1}\sum_{h=1}^H V_h^2$ et $B_2 = L^{-1}\sum_{h=1}^H V_h^4$. $\overline{\text{Var}(V_{wh}^h)}$.

Supposons que les conditions (C.1), (C.2) et (C.3) sont vérifiées. En outre, définissons $F_{s1}^{Ld_{s1}}$ et $F_{s2}^{Ld_{s2}}$, comme étant les premiers termes des développements en série de Taylor de $L^{-1}d_{s1} - \mu_{B_1}\mu_{A_1}^2$ et $L^{-1}d_{s2} - \mu_{B_2}\mu_{A_2}^2$, respectivement. En outre, rappelons que, si D suit la distribution d'une variable aléatoire khi-deux à d degrés de liberté, alors $V(D) = 2d$, $E(D^2) = d(d+2)$ et $V(D^2) = 8d(d+2)(d+3)$. Les composantes correspondantes de $\text{Var}(F_{s1}^{Ld_{s1}})$ et $\text{Var}(F_{s2}^{Ld_{s2}})$ dans (A.1) sont alors

$$= \frac{4}{\mu_{A_1}^2} (m_0 - 1) \mu_{A_2}^2 \text{Var}(A_2) - \frac{9}{\mu_{A_2}^2} (m_0 - 1) \mu_{A_1}^2 \text{Cov}(A_1, A_2) + \frac{4}{\mu_{A_1}^4} (m_0 - 1) \mu_{A_2}^4 \text{Var}(A_2)$$

$$- \frac{9}{\mu_{A_2}^3} (m_0 - 1) \text{Cov}(A_2, B_2) + \frac{4}{\mu_{A_1}^4} (m_0 - 1) \mu_{A_2}^4 \text{Var}(B_2) + \frac{27}{\mu_{A_2}^4} (m_0 + 1) \mu_{A_1}^4$$

$$= \frac{1}{1} (m_0 - 1) a - \frac{9}{1} (m_0 - 1) b + \frac{4}{1} (m_0 - 1) c \quad (\text{A.6})$$

$$\text{Var}(F_{s1}^{Ld_{s1}})$$

La substitution de (A.5) dans (A.1) donne

$$\text{Cov}(A_1, B_1) = 3(m_0 - 1)\beta_1^2 \text{Cov}(A_2, B_2). \quad (\text{A.5})$$

et

$$\text{Var}(B_1) = 12(m_0 + 1)^{-1} \mu_{A_1}^2 \text{Var}(B_2)$$

$$\text{Var}(A_1) = (m_0 - 1) \beta_1^2 \text{Var}(A_2)$$

$$\mu_{B_1} = 3\beta_1^2 \mu_{B_2}$$

Sous la condition (1.6), $\mu_{A_1} = \beta_1 \mu_{A_2}$,

$$L^{-1}d_{s2}^{ws} = L^{-1}d_{s2}^{ws}. \quad (\text{A.4})$$

et

$$L^{-1}d_{s1}^{ms} = L^{-1}(3L + 14)^{-1}(9L)d_{s1}^{s1} \quad (\text{A.3})$$

Puisque nous supposons que $n_h = 2$ et $m_h = m_0$ pour tout $h = 1, 2, \dots, L$, nous avons

$$\text{Cov}(A_2, B_2) = 4(m_0 - 1)^{-1} L^{-2} \sum_{h=1}^H V_h^3 \quad (\text{A.2})$$

et

$$\text{Cov}(A_1, B_1) = 12L^{-2} \sum_{h=1}^H V_h^3$$

$$\text{Var}(B_2) = 8(m_0 - 1)^{-2} (m_0 + 1) L^{-2} \sum_{h=1}^H V_h^4$$

$$\text{Var}(B_1) = 96L^{-2} \sum_{h=1}^H V_h^4$$

$$\text{Var}(A_2) = 2(m_0 - 1)^{-1} L^{-2} \sum_{h=1}^H V_h^{wh}$$

$$\text{Var}(A_1) = 2L^{-2} \sum_{h=1}^H V_h^2$$

5.3 Estimations des coefficients et du nombre de degrés de liberté

simulations pour t_1 sous chacun des cas 1 à 4. La colonne 8 du tableau 5.1 donne les valeurs p résultantes pour un test unilatéral. Les deux dernières colonnes du tableau 5.1 donnent les estimateurs du nombre de degrés de liberté d_{ms}^{HAR3} et d_{ms}^{BMPWT} . Pour HAR3 et BMPWT, d_{ms}^{HAR3} et d_{ms}^{BMPWT} sont produits des valeurs considérablement plus grandes que d_{ms}^{HAR3} .

6. Discussion

Nous avons considéré dans le présent article l'estimation d'un nombre de degrés de liberté d utilisé pour quantifier la variabilité d'un estimateur classique de variance fondé sur le plan $V(Y)$. Le problème fondamental est que, sous un plan comportant des variances au niveau de la strate hétérogènes et de petits nombres d'unités primaires d'échantillonnage par strate, l'estimateur de type Satterthwaite d_{ms}^{HAR3} peut donner de mauvais résultats. Nous avons élaboré un estimateur alternatif d_{ms}^{HAR3} basé sur les estimateurs de variance à l'intérieur des unités primaires d'échantillonnage V^{HAR3} . Cet estimateur alternatif est une solution d'une équation d'estimation sans biais (1.1) pour d , pour autant que la condition de proportionnalité (1.6) soit satisfait. En outre, la variance de la distribution approximative de d_{ms}^{HAR3} est plus petite que celle de d_{ms}^{HAR3} , à condition que le nombre d'unités secondaires d'échantillonnage sélectionnées dans chaque unité primaire soit grand, au sens défini par le résultat 2.2.

À la section 3, nous avons élaboré des méthodes à erreurs sur les variables pour tester l'adéquation de la condition de proportionnalité (1.6) et avons proposé certains diagnostics connexes. L'étude par simulation décrite à la section 4, conjuguée à l'analyse des données exposées à la section 5, indique que, sous un degré modéré d'hétérogénéité, d_{ms}^{HAR3} peut donner de meilleurs résultats que d_{ms}^{HAR3} en ce qui concerne les propriétés distributionnelles de ces estimateurs de d , de même que les taux de couverture et les largeurs des intervalles de confiance connexes pour les totaux de population X . Cependant, comme le laisse entendre la théorie classique des grands échantillons, ces estimateurs n'ont ni l'un ni l'autre de bonnes propriétés sous des conditions de forte hétérogénéité.

En principe, nous pourrions envisager l'utilisation des estimateurs à erreurs sur les variables ($\beta_0, \beta_1, \hat{\sigma}^{HAR3}$), ainsi que des estimateurs V^H et V^{HAR3} pour construire un estimateur alternatif de d qui serait convergent sous le modèle à erreurs sur les variables général (3.1)-(3.2) et ne nécessiterait pas la contrainte (1.6). Toutefois, les résultats des simulations présentés dans Jang (1996) indiquent que l'estimateur résultant d_{ms}^{HAR3} , disons, ne donne pas de bons résultats sous les conditions de simulation utilisées à la section 5.

Comme nos données étaient compatibles avec $\hat{\sigma}^{HAR3} = 0$ pour les quatre cas, nous avons utilisé les méthodes de Fuller (1987, page 124) pour produire les estimations de β_0 et de β_1 appropriées pour un modèle (3.1)-(3.2) sans erreur d'équation ; des renseignements détaillés peuvent être obtenus auprès des auteurs. Le tableau 5.1 donne aussi les résultats des estimations des coefficients β_0 et β_1 , et de leurs erreurs-types, $c_1(\beta_0)$ et $c_1(\beta_1)$. Rappelons, comme il est mentionné à la section 3.1, que sous le modèle (3.1)-(3.2), si $\beta_0 = 0$ et $\beta_1 \neq 0$, chaque variance de strate V^H est un multiple constant de la variance intra-UPE V^{HAR3} et d_{ms}^{HAR3} pourrait être un estimateur approprié de d . À la section 5.2, nous avons déjà pris en considération la condition $\hat{\sigma}^{HAR3} = 0$. Pour tester l'hypothèse nulle $H_0 : \beta_0 = 0$, nous utilisons la statistique de test $t_0 = \hat{\beta}_0 / c_1(\beta_0)$. Dans certains travaux pratiques portant sur les erreurs sur les variables, les quantités telles que t_0 sont comparées à une distribution de référence normale ou t standard. Cependant, les simulations basées sur les quatre cas décrits à la section 4.1 ont indiqué que la distribution nulle de t_0 s'écarterait considérablement des distributions de référence habituelles. Cela tient aux distributions très asymétriques des variables de réponse V^H utilisées dans la régression à erreurs sur les variables standard pour élaborer une distribution de référence fondée sur la simulation pour t_0 . La colonne 7 du tableau 5.1 donne la valeur p pour un test unilatéral à gauche. (En raison d'estimations ponctuelles négatives pour β_0 , nous avons choisi de présenter les valeurs p pour le test unilatéral à gauche ici. Dans d'autres cas, il pourrait être intéressant de présenter les valeurs p pour le test unilatéral à droite ou le test bilatéral pour β_0 .) Il existe des preuves convaincantes qu'il faut rejeter $H_0 : \beta_0 = 0$ pour la variable HDRESULT, et des preuves modérées qu'il faut rejeter $H_0 : \beta_0 = 0$ pour TCRESULT. Donc, il n'est peut-être pas approprié d'utiliser d_{ms}^{HAR3} pour ces deux variables. Maintenant, considérons le coefficient de pente β_1 et supposons que $\sigma^{HAR3} = 0$ de sorte que $\hat{\sigma}^{HAR3} = 0$ avec la probabilité de un. Alors, les expressions (1.5) et (3.1), ainsi que la non-négativité de V^{HAR3} impliquent que $0 \leq V^{HAR3} = V^H - V^{HAR3} = \beta_0 + (\beta_1 - 1)V^{HAR3}$. Conséquemment, si $\beta_0 = 0$, alors $\beta_1 \geq 1$ et $\beta_1 = 1$ est équivalent à $V^H = V^{HAR3}$. Cette dernière condition présente un intérêt pratique, parce que certains auteurs ont signalé des cas où V^{HAR3} est petite comparativement à V^H ou de manière équivalente, $V^H = V^{HAR3}$. Voir, par exemple, Wolter (1985, page 46). Pour tester $H_0 : \beta_1 = 1$ contre l'hypothèse alternative unilatérale $H_1 : \beta_1 < 1$, nous utilisons la statistique $t_1 = (\beta_1 - 1) / c_1(\beta_1)$. Pour des raisons similaires à celles énoncées pour t_0 , nous avons élaboré des distributions de référence fondées sur les

Figure 5.1 Tracé de F_{wh} en fonction de V_h pour les mexicaino-américains de 20 à 29 ans, variable = BMPWT

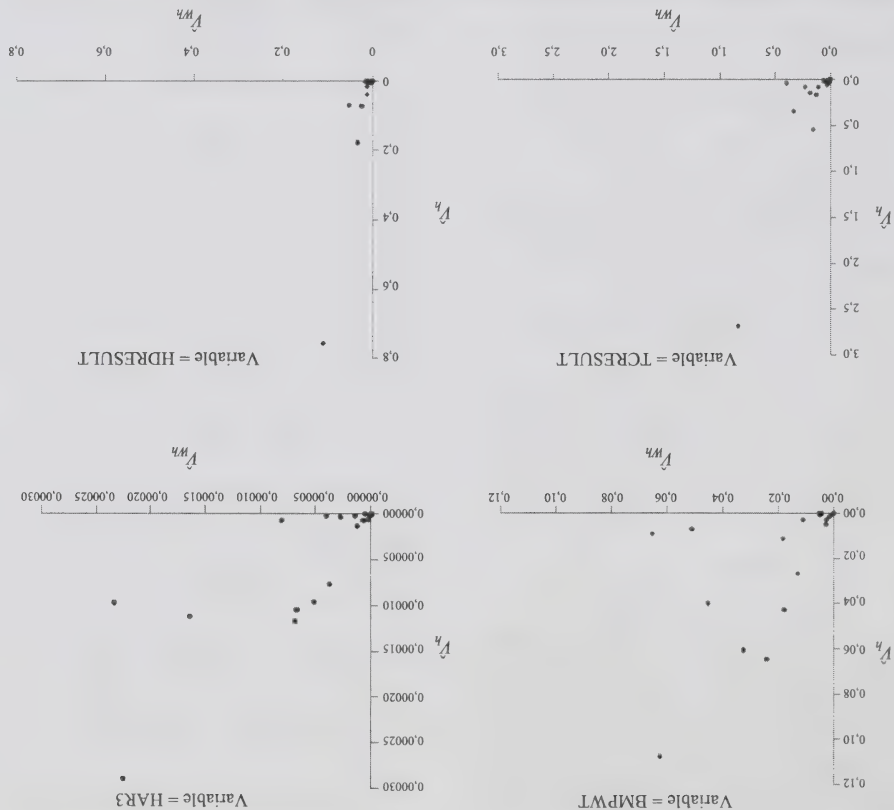


Tableau 5.2 Moyennes et quantiles de $f_{qq} = \sigma^{-1} \hat{\sigma}_{qq} (\beta_0 = 0)$

Cas	$^1M(f_{qq})$	$E-T.(f_{qq})$	$^2Q(0.01)$	$Q(0.05)$	$Q(0.10)$	$Q(0.25)$	$Q(0.50)$	$Q(0.75)$	$Q(0.90)$	$Q(0.95)$	$Q(0.99)$
1	-0,50	0,66	-1,71	-1,30	-1,15	-0,99	-0,79	0,16	0,54	0,60	0,65
2	-0,48	0,68	-1,72	-1,32	-1,16	-0,99	-0,76	0,23	0,57	0,62	0,66
3	-0,19	0,42	-1,01	-0,84	-0,74	-0,53	-0,20	0,17	0,38	0,46	0,55
4	-0,20	0,39	-1,00	-0,82	-0,72	-0,51	-0,20	0,11	0,34	0,44	0,56

¹ M désigne la moyenne des estimations calculées sur l'ensemble des 10 000 répliques.
² Q(.) indique le quantile de l'estimateur calculé sur l'ensemble des 10 000 répliques.

Tableau 5.3 Moyennes et quantiles de $f_{qq} = \sigma^{-1} \hat{\sigma}_{qq}$

Cas	$^1M(f_{qq})$	$E-T.(f_{qq})$	$^2Q(0.01)$	$Q(0.05)$	$Q(0.10)$	$Q(0.25)$	$Q(0.50)$	$Q(0.75)$	$Q(0.90)$	$Q(0.95)$	$Q(0.99)$
1	-0,56	0,62	-1,85	-1,34	-1,17	-1,00	-0,80	0,05	0,38	0,44	0,52
2	-0,56	0,62	-1,91	-1,37	-1,18	-1,00	-0,78	0,06	0,35	0,42	0,50
3	-0,24	0,42	-1,16	-0,90	-0,79	-0,57	-0,22	0,12	0,29	0,36	0,45
4	-0,24	0,38	-1,09	-0,87	-0,75	-0,53	-0,22	0,06	0,25	0,33	0,44

¹ M désigne la moyenne des estimations calculées sur l'ensemble des 10 000 répliques.
² Q(.) indique le quantile de l'estimateur calculé sur l'ensemble des 10 000 répliques.

Tableau 4.3b
Probabilités de non-ouverture simulées, et moyennes et quantiles des multiplicateurs / pour les intervalles de confiance à 95 % nominaux : variances réelles négatives, cas 1 et 2

Cas	Est.	$1 - \hat{\alpha}$	${}^2M(2\hat{\alpha})$	$E.-T. (2\hat{\alpha})$	${}^3Q(0,05)$	$Q(0,25)$	$Q(0,50)$	$Q(0,75)$	$Q(0,95)$
1	$\hat{\alpha}_{ms}$	0,0552	4,62	0,36	4,26	4,38	4,52	4,75	5,37
	$\hat{\alpha}_{hs}$	0,0428	4,83	0,16	4,64	4,72	4,80	4,90	5,13
2	$\hat{\alpha}_{ms}$	0,0567	4,66	0,36	4,29	4,41	4,55	4,78	5,41
	$\hat{\alpha}_{hs}$	0,0466	4,87	0,21	4,64	4,72	4,83	4,97	5,28
	$n - L$	0,0788	4,15						
	$d_s\text{ réel}$	0,0443	4,89						

¹ $1 - \hat{\alpha}$ est la probabilité de non-ouverture simulée des intervalles de confiance calculée en utilisant le nombre estimé de degrés de liberté.
² ${}^2M(2\hat{\alpha}_{0,975})$ est la moyenne de deux fois la valeur du 97,5^e centile de la loi $Q(\cdot)$ indique le quantile de $2\tau_{0,975,\hat{d}}$ calculé sur la totalité des répliques.

Tableau 5.1
Estimations des paramètres du modèle, diagnostics du modèle et estimations du nombre de degrés de liberté pour les variables de la NHANES III (sous-groupes des mexicano-américains de 20 à 29 ans)

Variables	K_{xx}	β_0	$E.-T.(\beta_0)$	β_1	$E.-T.(\beta_1)$	Valeur p fondée sur la simulation pour $H_0 : \beta_0 = 0$ pour $H_0 : \beta_1 = 1$	Valeur p fondée sur la simulation	$\hat{\sigma}_{qq}$	F_{qq}	d_{ms}	d_{hs}
BMPWT	0,75	-0,0013	0,0039	1,135	0,5429	0,3815	0,3541	-0,000	-0,43	15,49	10,04
HAR3	0,75	-0,000009	0,000012	1,095	0,3991	0,4229	0,3400	0,000	-0,83	14,94	8,30
TCEFSULT	0,88	-0,146	0,0493	2,879	0,6252	0,0606	0,2259	-0,178	-0,77	5,88	6,59
HDRFSULT	0,90	-0,042	0,0098	6,650	0,9988	<0,0001	0,1506	-0,017	-0,91	5,45	5,93

5.2 Test ponctuel de $\sigma_{qq} = 0$ sous la condition (C.1)

Pour les quatre variables considérées au tableau 5.1, les estimations directes $\hat{\sigma}_{qq}$ de la variance de l'erreur d'équation (3.4) étaient négatives ou proches de zéro. Cela donne à penser que notre estimateur basé sur χ^2 de σ_{qq} donné à la section 3.1 pourrait être trop prudent ou que $\hat{\sigma}_{qq}$ est effectivement proche de zéro. Il semble donc que nous devons réexaminer l'hypothèse distributionnelle (C.1) dans l'exemple de la NHANES III. Pour cela, nous considérons la distribution simulée de $F_{qq}^{est} = \hat{\sigma}_{qq}^2 / \hat{\sigma}_{ee}^2$, où nous recourons à la division par $\hat{\sigma}_{ee}$ pour éviter les problèmes d'échelle. Les conditions et le plan de simulation sont les mêmes que ceux décrits à la section 4.1.

Le tableau 5.2 donne les résultats pour $\hat{\sigma}_{ee}$ d'après l'expression (3.5) ainsi que $\hat{\sigma}_{qq}$ calculé d'après l'expression (3.4) avec β_0 fixé égal à zéro et avec β_1 calculé d'après l'expression (3.3). Le tableau 5.2 contient la moyenne, l'écart-type et certains quantiles de la distribution simulée de F_{qq} pour les quatre variables. Le tableau 5.3 donne les quantités correspondantes pour F_{qq} calculé d'après $\hat{\sigma}_{qq}$ donné par l'expression (3.4) et avec β_0 et β_1 calculés d'après l'expression (3.3).

Les résultats présentés aux tableaux 5.2 et 5.3 mettent à un test ponctuel de vérification de $H_0 : \sigma_{qq} = 0$. Précisément, si le ratio observé F_{qq} se situe au-dessus du quantile supérieur 0,95 simulé, alors l'hypothèse que $\sigma_{qq} = 0$ pourrait être problématique. Inversement, un F_{qq} observé de valeur inférieure au quantile 0,05 simulé dans les tableaux 5.2 ou 5.3 pourrait indiquer que $\hat{\sigma}_{qq}$ est prudent, ou que d'autres éléments de la condition (C.1) sont violés.

D'après le tableau 5.1, les valeurs de F_{qq} pour les variables sont comprises entre -0,91 et -0,43. Sauf pour HDRFSULT, nous ne disposons d'aucune preuve convaincante d'une violation des hypothèses du modèle. Cependant, pour HDRFSULT, le ratio $F_{qq} = -0,91$ se situe entre les valeurs des quantiles 0,01 et 0,05 présentées dans les tableaux 5.2 et 5.3 pour le cas 4. En général, les valeurs de F_{qq} qui tombent au-delà des quantiles 0,95 ou 0,99 des tableaux 5.2 ou 5.3 seraient compatibles avec des valeurs de $\hat{\sigma}_{qq}$ supérieures à zéro. La valeur observée $F_{qq} = -0,91$ ne correspond pas nécessairement à $\hat{\sigma}_{qq} > 0$, mais indique la violation d'une ou de plusieurs conditions dans (C.1) à (C.4).

4.2 Taux de couverture des intervalles de confiance

fondés sur t

Pour les quatre cas spécifiés, le tableau 4.2 donne les probabilités simulées de non-couverture obtenues pour les intervalles de confiance fondés sur t pour la moyenne de population \bar{Y} obtenue en utilisant l'estimateur \hat{d} correspondant. Dans les cas fortement hétérogènes (cas 3 et 4), aucune mesure du nombre de degrés de liberté (pas même le d réel) ne donne des intervalles de confiance dont le taux de couverture correspond au taux $1 - \alpha$. Autrement dit, dans les cas extrêmes, l'approche générale de Satterthwaite peut poser des problèmes pour la construction des intervalles de confiance, que l'on utilise \hat{d} , \hat{d}_{ms} ou \hat{d}_{ws} pour déterminer le multiplicateur t .

Dans les cas 1 et 2, les valeurs V_h présentent une hétérogénéité moins importante que dans les cas 3 et 4. Le tableau 4.2 montre que les probabilités de couverture simulées avec la valeur réelle d pour ces deux cas sont légèrement supérieures à 0,95. Cette couverture excédentaire pourrait être attribuable au fait que l'estimateur de variance $V(Y)$ n'est pas distribué exactement comme un multiple d'une variable aléatoire χ^2_d , à cause de l'hétérogénéité des V_h . L'utilisation du nombre classique de degrés de liberté $n - L$ ou de l'estimateur modifié \hat{d}_{ms} produit des intervalles de confiance dont le taux de couverture est inférieur au niveau nominal de 95 %. Par ailleurs, l'utilisation de notre nombre fondé sur des données auxiliaires \hat{d}_{ws} donne des taux de couverture basés sur la simulation proches de ce niveau nominal.

Les tableaux 4.3a et 4.3b présentent les distributions empiriques de \hat{d} et de $2t_{\hat{d}}$ pour les estimateurs \hat{d}_{ms} et

5. Application à une enquête sur la santé

5.1 Vérifications préliminaires du modèle

Nous avons appliqué les méthodes que nous proposons aux données de la NHANES III décrites à la section 1. Il est important que nous vérifions les hypothèses de modélisation avant d'appliquer les mesures proposées de stabilité. En premier lieu, pour la sous-population de mexicano-américains de R_{xx} pour les quatre variables ayant une valeur de R_{xx} supérieure à 0,7.

En deuxième lieu, la figure 5.1 présente les diagrammes de dispersion de V_h en fonction de V_{ws} pour les quatre variables en utilisant les mêmes échelles pour les axes horizontal et vertical. Ces diagrammes montrent qu'une relation linéaire est plausible pour les variables correspondant, même si la relation n'est pas parfaite et qu'il existe certaines valeurs aberrantes. Par conséquent, la méthode fondée sur des données auxiliaires élaborées aux sections 2 et 3 pourraient convenir pour ces quatre variables.

Cas	d réel	Est.	\hat{d} moyen	$E-T$ (\hat{d})	$Q(0,05)$	$Q(0,25)$	$Q(0,50)$	$Q(0,75)$	$Q(0,95)$
Moyennes et quantiles des estimateurs du nombre de degrés de liberté \hat{d}_{ms} et \hat{d}_{ws} : cas 1 et 2									
1	6,26	\hat{d}_{ms} 9,33 \hat{d}_{ws} 6,52	3,33	4,45	6,86	9,01	11,41	15,30	7,78
2	6,04	\hat{d}_{ms} 8,87 \hat{d}_{ws} 6,34	2,95	4,35	6,69	8,72	10,97	13,99	7,80
¹ Moyen désigne la moyenne des estimations calculées sur l'ensemble des 10 000 répliques.									
² $Q(\cdot)$ indique le quantile de l'estimateur calculé sur l'ensemble des 10 000 répliques.									

d_s réel	6,26	6,04	6,04	6,04	6,04	6,04	6,04	6,04	6,04
Non-couverture avec \hat{d}_{ms}	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428
Non-couverture avec \hat{d}_{ws}	0,0552	0,0552	0,0552	0,0552	0,0552	0,0552	0,0552	0,0552	0,0552
Non-couverture avec t_{n-L}	0,0744	0,0744	0,0744	0,0744	0,0744	0,0744	0,0744	0,0744	0,0744
Non-couverture avec t_{d_s}	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428	0,0428
d_s réel	6,26	6,04	6,04	6,04	6,04	6,04	6,04	6,04	6,04
Cas 1	Cas 2	Cas 3	Cas 4	Cas 4	Cas 4	Cas 4	Cas 4	Cas 4	Cas 4

Strate	Cas 1	Cas 2	Cas 3	Cas 4
1	0,00E+00	0,00E+00	0,00E+00	0,00E+00
2	0,00E+00	0,00E+00	0,00E+00	0,00E+00
3	1,56E-04	7,67E-05	1,45E-02	1,76E-02
4	2,01E-04	3,57E-06	5,60E-02	4,55E-03
5	2,82E-04	4,88E-07	1,54E-03	2,91E-03
6	4,36E-04	0,00E+00	3,73E-03	8,60E-04
7	7,30E-04	2,14E-06	1,69E-02	1,13E-05
8	8,80E-04	1,30E-05	2,72E-02	1,40E-03
9	1,65E-03	1,16E-06	9,24E-03	1,35E-04
10	1,70E-03	9,46E-07	2,24E-03	1,77E-03
11	2,73E-03	0,00E+00	2,54E-04	1,32E-03
12	2,91E-03	5,40E-06	2,75E-02	6,40E-03
13	4,95E-03	3,73E-07	1,15E-02	5,38E-03
14	7,25E-03	2,90E-04	3,75E-02	6,97E-02
15	9,06E-03	9,81E-05	3,46E-01	7,58E-01
16	1,14E-02	7,47E-06	1,54E-02	4,75E-03
17	2,69E-02	9,65E-05	7,99E-02	1,01E-03
18	4,00E-02	1,12E-04	1,44E-01	1,77E-01
19	4,27E-02	2,68E-06	8,59E-02	3,88E-02
20	6,05E-02	7,57E-06	2,68E+00	7,18E-02
21	6,45E-02	1,17E-04	1,65E-01	4,52E-04
22	1,08E-01	1,05E-04	5,41E-01	1,98E-03

Tableau 4.1
Variances « réelles » V_h utilisées dans les études en simulation

Nous recourons maintenant à une étude par simulation pour évaluer les propriétés de nos estimateurs du nombre de degrés de liberté et les variables connexes, sous des conditions de taille d'échantillon modérée. Nous établissons la procédure de simulation comme il suit.

Nous considérons quatre ensembles de valeurs de V_h provenant de l'exemple de la NHANES III pour la sous-population mexicano-américaine présentée à la section 1.1. Ces quatre ensembles de V_h sont les valeurs estimées V_h

4.1 Plan de l'étude

4. Une étude par simulation

Le travail de la présente section est basé sur l'hypothèse que $\sigma^{qq} > 0$. Il est possible d'élaborer des diagnostics connexes applicables au cas de l'absence d'erreur d'équation, c'est-à-dire $\sigma^{qq} = 0$; des renseignements détaillés peuvent être obtenus auprès des auteurs.

ces diagnostics de manière plus approfondie.

numérique présentée aux sections 4 et 5, nous examinerons grandes dans l'information auxiliaire V_{wh} . Dans le travail résultats, à cause des erreurs d'échantillonnage relativement grandes dans l'information auxiliaire V_{wh} . Dans le travail

pour les variables BMPWT, HAR3, TCRSULT et HDRSULT, respectivement, et sont énumérés au tableau 4.1. Pour chaque cas, nous avons utilisé $(\beta_0, \beta_1) = (0, 1)$ et $\sigma^{qq} = 0$, en harmonie avec les résultats de la section 3 et, donc, $V_{wh} = V_h$. Puis, pour chaque $h = 1, \dots, L$, nous avons obtenu 10 000 réalisations des estimateurs initiaux $(\hat{V}_{h1}, \hat{V}_{h2}, \hat{V}_{wh1}, \hat{V}_{wh2})$ en supposant que les \hat{X}_{hi} sont distribués comme une variable aléatoire normale de moyenne nulle et de variance $2^{-1}V_h$, que $V^{-1}(m_{hi} - 1)V_{wh}$ est degrés de liberté, où $m_{hi} = 11$ pour tout h et i , et que les \hat{X}_{hi} et \hat{V}_{wh} sont mutuellement indépendants. Notons que, dans nos données provenant de la NHANES III, le nombre moyen d'unités secondaires dans chaque UPE i dans la strate h est de l'ordre de 11. Pour chaque réplique, nous avons calculé $V_h = (\hat{X}_{h1} - \hat{X}_{h2})^2$ et $V_{wh} = 2^{-1}(\hat{V}_{wh1} + \hat{V}_{wh2})$, puis nous avons effectué une régression à erreurs sur les variables de V_h sur V_{wh} avec la variance de l'erreur de mesure $\hat{\sigma}_{unh}^2 = \text{Var}(V_{wh})$ donné par la formule (2.2). Nous avons obtenu ainsi les estimateurs des coefficients $(\hat{\beta}_0, \hat{\beta}_1)$, ainsi que les estimateurs du nombre de degrés de liberté \hat{d}_{ms} et \hat{d}_{ws} .

Jang et Elling : Utilisation des variances à l'intérieur des UPE pour évaluer la stabilité d'un estimateur

mentionnées. Pour les quatre cas présentés au tableau 1.1 et étudiés plus en profondeur à la section 4 qui suit, $g(a,b,c)$ est égal à 4.7, 4.3, 4.6 et 4.8, respectivement, alors que, selon l'application de la NHANES III, la moyenne des m_h valeurs était approximativement égale à 22. En outre, nous

Supposons que V^w/V^w possède les mêmes moments que F/f , où F suit une loi du khi-deux à f degrés de liberté. Alors, $f = \infty$ correspond au cas où $V^w = F^w$ pour tout h , qui correspond au cas où la valeur réelle de d dans (1.1) est égale à la valeur habituelle $n - L$.

3. Test de la condition de proportionnalité

3.1 Un modèle à erreurs sur les variables pour V_h et V^w_h

L'élaboration de l'estimateur de rechange d^{ws} à la section 1 et l'évaluation de ses propriétés à la section 2 dépendaient fortement de la condition de proportionnalité (1.6). Il est possible de tester l'adéquation de cette condition comme il suit. Premièrement, notons que la condition (1.6) est un cas particulier du modèle qui suit.

$$(C.4) \quad \text{Pour tout } h = 1, 2, \dots, L,$$

$$(3.1) \quad V_h = \beta_0 + \beta_1 V^w_h + q_h$$

où β_0 et β_1 sont des constantes, et q_h est une erreur d'équation de moyenne nulle et de variance σ^{qgh} .

Deuxièmement, rappelons que V_h et V^w_h sont des quantités inconnues pour lesquelles nous avons les estimateurs sans biais, \hat{V}_h et \hat{V}^w_h , respectivement. En utilisant la notation des modèles à erreurs sur les variables utilisée dans Fuller (1987), définissons les erreurs d'estimation

$$(3.2) \quad e_h = \hat{V}_h - V_h \quad \text{et} \quad u_h = \hat{V}^w_h - V^w_h.$$

Sous les conditions (C.1) et (C.2), le vecteur $(e_h, u_h)'$ possède une distribution dont le vecteur moyen est égal à $(0, 0)'$ et la matrice de variance-covariance est égale à $\text{diag}(\sigma^{ech}, \sigma^{uuh})$, où $\sigma^{ech} = (m_h - 1)^{-1} 2V^2_h$ et $\sigma^{uuh} = (m_h - 1)^{-1} 2V^2_{wh}$. Sous la condition supplémentaire (C.3), ces termes de variance se simplifient en $\sigma^{ech} = 2V^2_h$ et $\sigma^{uuh} = (m_h - 1)^{-1} 2V^2_{wh}$. Les expressions (3.1) et (3.2) définissent un modèle de régression à erreurs sur les variables contenant des variances d'erreur de mesure hétérogènes et des erreurs ne suivant pas une loi normale. En outre, $\text{Var}(V^w_h)$ défini dans l'expression (2.2) est un estimateur sans biais de σ^{uuh} donc de l'information d'identification pour les paramètres β_0, β_1 et σ^{qgh} dans le modèle (3.1) - (3.2). Une application

directe de Fuller (1987, pages 187 à 189) avec poids égaux donne alors les estimateurs convergents (pour L croissant),

$$(3.3) \quad \hat{\beta}_0 = L^{-1} \sum_{h=1}^L V^w_h - \hat{\beta}_1 \hat{V}^w_h, \quad \hat{\beta}_1 = \left[\sum_{h=1}^L (V^w_h - \hat{V}^w_h)^2 - \hat{\sigma}^{uuh} \right]^{-1} \left[\sum_{h=1}^L (V^w_h - \hat{V}^w_h) V^w_h - \hat{V}^w_h \hat{V}^w_h \right],$$

et

$$\hat{\sigma}^{qgh} = \max \left[0, L^{-1} \sum_{h=1}^L (m_h - 1)^{-1} \{ (L - 2)^{-1} L (V^w_h - \hat{\beta}_0 - \hat{\beta}_1 V^w_h)^2 - (\hat{\sigma}^{ech} + \hat{\beta}_1^2 \hat{\sigma}^{uuh}) \} \right],$$

où

$$(3.4) \quad \hat{\sigma}^{uuh} = \sum_{h=1}^L \widehat{\text{Var}}(V^w_h), \quad \hat{V}^w_h = L^{-1} \sum_{h=1}^L V^w_h,$$

et

$$(3.5) \quad \hat{\sigma}^{ech} = 2(m_h + 1)^{-1} V^2_h$$

d'après la condition (C.1). De plus, l'application directe de Fuller (1987, page 188) mène aux estimateurs de variance $\hat{V}(\beta_0)$ et $\hat{V}(\beta_1)$, disons : des renseignements détaillés peuvent être obtenus auprès des auteurs.

3.2 Deux diagnostics connexes

Conformément à la condition (C.4), l'estimateur proposé d^{ws} est destiné à être appliqué aux cas où les V^w_h fournissent de l'information utile sur les grands variances globales au niveau de la strate V_h . Pour repérer ce cas, un diagnostic simple est le ratio $\{ \hat{V}(V_h) \}^{-1} \{ \hat{\beta}_1 \hat{V}(V^w_h) + \hat{\sigma}^{qgh} \}$, c'est-à-dire le ratio des estimateurs des variances des distributions approximatives de $V_h - V_h$ et de $\beta_1 V^w_h - V^w_h$, respectivement, sous le modèle (3.1) - (3.2). Si ce ratio est considérablement inférieur à l'unité, l'utilisation de d^{ws} pourrait être indiquée. De surcroît, la performance de l'estimateur d^{ws} dépend fortement de la grandeur de $\hat{\sigma}^{uuh}$ relativement à la variabilité des variances intra-UPF réelles V^w_h . Définissons un estimateur du ratio de fiabilité (Fuller 1987, page 3)

$$k^{xx} = \max \left\{ 0, \left[\sum_{h=1}^L (V^w_h - \hat{V}^w_h)^2 \right]^{-1} \left[\sum_{h=1}^L (V^w_h - \hat{V}^w_h)^2 - \hat{\sigma}^{uuh} \right] \right\}.$$

Les valeurs de k^{xx} sont comprises entre 0 et 1, et les valeurs de k^{xx} proches de l'unité indiquent que les erreurs dans l'estimation des variances intra-UPF sont relativement faibles. Inversement, de faibles valeurs de k^{xx} (par exemple, $k^{xx} < 0,7$) pourraient indiquer que les méthodes

V_h sont des fonctions de la moyenne d'échantillon des $P_{hij}^{V_h}$ sur les UPE dans la strate h . En outre, les estimateurs $V_h^{W_h}$ sont des fonctions des variances d'échantillon des $P_{hij}^{V_h}$ dans l'UPE (h, i). Donc, dans le cas de variables X pour lesquelles les $P_{hij}^{V_h}$ suivent approximativement une loi normale dans la strate h , les estimateurs V_h et $V_h^{W_h}$ sont approximativement indépendants.

2.2 Propriétés de d^{ws}

Dans la suite de l'article, nous utiliserons dans l'estimateur d^{ws} défini par l'expression (1.8) l'estimateur

$\text{Var}(V_h^{W_h})$ tel qu'il est défini dans l'expression (2.2). En outre, nous utiliserons plusieurs résultats asymptotiques. Ces résultats s'appuieront sur la condition que le nombre de strates L augmentent, tandis que les tailles des échantillons d'UPE et d'USE au niveau de la strate n_h et m_h peuvent demeurer petites. Cette condition est en harmonie avec de nombreux plans à plusieurs degrés pour lesquels, en pratique, $n_h = 2$ et les valeurs de m_h sont modérées. Voir, par exemple, Krewski et Rao (1981) pour un exposé détaillé des résultats asymptotiques pour une grande valeur de L . La preuve du résultat 2.1 étant un exercice de routine, nous l'omettons ici.

Résultat 2.1. Supposons que $E(V_h^{W_h}) = O(1)$ pour $r = 1, 2, 3, 4$ et définissons

$$\bar{V}^W = L^{-1} \sum_{h=1}^H V_h^{W_h}$$

et

$$\hat{V}^{W(2)} = L^{-1} \sum_{h=1}^H (n_h^{-1} - 1)^{-1} \{V_h^{W_h(2)} - \widehat{\text{Var}}(V_h^{W_h})\}.$$

Alors \bar{V}^W et $\hat{V}^{W(2)}$ sont des estimateurs convergents de V^W et $L^{-1} \sum_{h=1}^H (n_h^{-1} - 1)^{-1} V_h^{W_h(2)}$ respectivement. En outre, $L^{-1} d^{ws}$ est un estimateur convergent de $L^{-1} d^{ws}$. Nous avons suggéré à la section 1 que, dans certains cas, l'estimateur d^{ws} fondé sur des données auxiliaires pourrait être plus stable que l'estimateur de Satterthwaite modifié \hat{d}^{ms} . Pour examiner cette idée, nous allons comparer les variances de d^{ws} et \hat{d}^{ms} sous la condition (C.1) et les hypothèses supplémentaires qui suivent.

(C.2) Pour $h = 1, 2, \dots, L$, les $V_h^{W_h} (m_h - 1)^{-1} V_h^{W_h}$ sont distribués comme des variables aléatoires khidépendantes à $m_h - 1$ degrés de liberté, respectivement, où m_h est le nombre d'USE dans la strate h , et sont mutuellement indépendants de $V_h^{W_h}$.

(C.3) Pour tout $h = 1, 2, \dots, L$, $n_h = 2$, et $m_h = m_0$ pour un entier positif fixé $m_0 \geq 2$.

i) les variances du premier terme des développements en série de Taylor de $L^{-1} (d^{ws} - d)$ et de $L^{-1} (d^{ms} - d)$ sont, respectivement,

$$V^{LW} = a - b + c$$

et

ii) pour tout $m_0 \geq \lim_{L \rightarrow \infty} g(a, b, c)$, $\lim_{L \rightarrow \infty} V^{LW} \geq \lim_{L \rightarrow \infty} V^{Lm}$, où

$$g(a, b, c) = \{2(3a - 3b + 4c)\}^{-1} \left\{ 1 + c + \sqrt{144a^2 + 144b^2 + 153c^2 - 288ab + 216ac - 216bc} \right\}.$$

iii) pour $m_0 \geq 10$, $\lim_{L \rightarrow \infty} V^{Lm} \geq \lim_{L \rightarrow \infty} V^{LW}$ indépendamment des valeurs des moments limites $\lim_{L \rightarrow \infty} (\mu_{k_j}^{LW})$, $\lim_{L \rightarrow \infty} (\mu_{k_j}^{Lm})$, $L^{-1} \sum_{h=1}^H \mu_{k_j}^{W_h}$, $L^{-1} \sum_{h=1}^H \mu_{k_j}^{W_h}$.

Le résultat 2.2 indique que, si L est grand, d^{ws} pourrait être préférable à \hat{d}^{ms} , pour autant que 1) la condition de proportionnalité (1.6) soit satisfaite et que 2) la taille de l'échantillon d'unités secondaires m_0 soit supérieure à la borne inférieure donnée par $g(a, b, c)$ (faisant en sorte que les variances des $V_h^{W_h}$ soient relativement faibles). Ce résultat motive l'emploi des variances intra-UPE pour évaluer la stabilité des estimateurs de variance utilisés dans les enquêtes, tout spécialement sous des plans d'échantillonnage dans lesquels les nombres d'UPE par strate sont faibles. Pour une discussion supplémentaire de ce point et de certains diagnostics précis en vue de vérifier la stabilité de $V_h^{W_h}$, voir Ellingre et Jang (1996) et les références

1.3 Utilisation de données auxiliaires au niveau de la strate

Dans le cas où l'hétérogénéité des termes V_h est

modérée, des travaux de simulation menés par Jang (1996) ont indiqué que d_{ms} donne d'assez bons résultats. Cependant, si l'hétérogénéité des variances de strate est importante (c'est-à-dire si $L^{-1}d$ est relativement petit), d_{ms}

peut être insatisfaisant. Le problème fondamental tient au fait que, si les valeurs de n_h sont relativement faibles, à eux seuls, les estimateurs V_h ne fournissent pas suffisamment d'information concernant les grands variances relatives des variances réelles au niveau de la strate V_h . Dans certains cas, un estimateur de variance fondé sur des données auxiliaires devrait, en principe, être plus stable que l'estimateur habituel fondé sur le plan; voir, par exemple, Isaki (1983). De même, des sources auxiliaires d'information peuvent être utilisées pour évaluer les grands variances relatives des variances V_h .

Dans la suite de l'article, nous nous concentrons sur l'information auxiliaire fournie par les relations entre les variances globales au niveau de la strate V_h et les variances intra-UPF connexes. Rappelons, suivant Wolter (1985, page 41), la décomposition

$$\text{Var}(Y_h) = V_{Bh} + V_{wh}, \quad (1.5)$$

où $V_{Bh} = \text{Var}\{\sum_{i=1}^{N_h} (n_h p_{hi})^{-1} Y_{hi}\}$ est la variance inter-UPF, $V_{wh} = \sum_{i=1}^{N_h} (n_h p_{hi})^{-1} \sigma_{2hi}^2$ est la variance intra-UPF, $Y_{hi} = E(Y_{hi} | \text{UPF } i, \text{ strate } h)$ et $\sigma_{2hi}^2 = \text{Var}(Y_{hi} | \text{UPF } i, \text{ strate } h)$. En outre, définissons $V^w = L^{-1} \sum_{h=1}^L V_{wh}$.

Les estimateurs de V^w peuvent fournir de l'information auxiliaire utile sur les grands variances relatives de V_h pour deux raisons. Premièrement, sous les plans où n_h est petit et n_{hi} est relativement grand, les estimateurs de variance intra-UPF V_{wh} peuvent être considérablement plus stables que V_h . Deuxièmement, dans certaines applications (par exemple, certains exemples de variance observées sont convergentes par rapport à un modèle sous lequel V_h est proportionnelle à V^w , c'est-à-dire

$$V_h = \beta_1 V^w \text{ pour tout } h = 1, \dots, L, \quad (1.6)$$

où β_1 est une constante fixée. La relation de proportionnalité (1.6) se manifestera si V_{Bh} et V_{wh} sont toutes deux proportionnelles à un facteur d'échelle commun, par exemple, $(\frac{V_h}{L})^\alpha$ pour une puissance donnée α . Sous la relation (1.6), l'expression (1.2) peut être réécrite sous la forme

$$d = \left\{ \sum_{h=1}^L (n_h - 1)^{-1} V_{wh}^2 \right\}^{-1} \left\{ \sum_{h=1}^L V_{wh}^2 \right\}^{-1} \quad (1.7)$$

2. Un estimateur fondé sur de l'information auxiliaire

2.1 Un estimateur de variance intra-UPF

Un estimateur simple de V^w est donné par

$$V^w_{wh} = n_h^{-2} \sum_{i=1}^{N_h} p_{hi}^{-2} \sigma_{2hi}^2, \quad (2.1)$$

où $\sigma_{2hi}^2 = n_{hi}^{-1} (n_h - 1)^{-1} \sum_{j=1}^{N_h} (p_{hj}^{-1} p_{hi}^{-1} y_{hj}^2 - \bar{y}_{hi}^2)$. Notons que σ_{2hi}^2 est approximativement sans biais pour σ_{2hi}^2 sous un plan d'échantillonnage avec remise à l'intérieur de l'UPF; dans la strate h ou sous un échantillonnage aléatoire simple sans remise et avec faible fraction d'échantillonnage, $f_{hi} = N_{hi}^{-1} n_{hi}$. La théorie classique de l'échantillonnage montre que V^w_{wh} est approximativement sans biais pour V^w_{wh} . Alors, un estimateur approximativement sans biais de $\text{Var}(V^w_{wh})$ est donné par

$$\widehat{\text{Var}}(V^w_{wh}) = n_{hi}^{-1} (n_h - 1)^{-1} \sum_{i=1}^{N_h} (V^w_{whi} - V^w_{wh})^2, \quad (2.2)$$

À la section 4, dans une étude en simulation, nous explorons les conditions sous lesquelles le nouvel estimateur proposé d_{ms} peut donner de meilleurs résultats que d_{ms} . Cette évaluation s'appuie à la fois sur l'estimation de d en tant que tel et sur les propriétés des intervalles de confiance pour Y . À la section 5, nous appliquons l'estimateur proposé à quatre variables de la NHANES III, en insistant sur les cas pour lesquels les différences entre les estimateurs proposés d_{ms} et d_{ms} ont un effet considérable en pratique sur l'évaluation de la stabilité de l'estimateur de variance $V(Y)$. Enfin, à la section 6, nous passons en revue les méthodes élaborées dans le présent article et envisageons certaines extensions possibles.

est un estimateur alternatif de d . À la section 2, nous examinons certaines propriétés de d_{ms} . À la section 3.1, nous utilisons des tests à erreurs sur les variables pour vérifier l'adéquation de la condition de proportionnalité (1.6). À la section 3.2, nous présentons deux diagnostics connexes pour la relation entre V_h et des variables auxiliaires, et pour la grandeur de l'erreur dans les variables auxiliaires observées V^w_{wh} .

$$d_{ms} = \left\{ \sum_{h=1}^L (n_h - 1)^{-1} [V^w_{wh} - \widehat{\text{Var}}(V^w_{wh})] \right\}^{-1} \left\{ \sum_{h=1}^L V^w_{wh} \right\}^{-1} \quad (1.8)$$

Par conséquent, étant donné un ensemble d'estimateurs de variance intra-UPF stables V^w_{wh} et d'estimateurs de variance de variance connexes $\widehat{\text{Var}}(V^w_{wh})$,

où $V^w_{whi} = n_{hi}^{-1} p_{hi}^{-2} \sigma_{2hi}^2$; voir, par exemple, Elling et Jang (1996) et les références mentionnées par ces auteurs. Notons que les estimateurs de variance globale au niveau de la strate

Tableau 1.1
Quatre variables de la NHANES III

Nom de la variable		Description
BMPWT	Poids (kg)	
HAR3	Fumez-vous des cigarettes à l'heure actuelle ? (0/1)	
TCRESULT	Cholestérol sérique total (mg/dL)	
HDRESULT	Cholestérol HDL (mg/dL)	

1.2 Stabilité des estimateurs de variance fondés sur le plan

Supposons que nous ayons une population partitionnée en L strates, avec N_h UPF dans la strate h pour $h = 1, 2, \dots, L$. Sous un plan d'échantillonnage à plusieurs degrés stratifié, nous sélectionnons n_h UPF avec remise et avec probabilité de sélection par tirage p_{hi} pour l'UPF i dans la strate h , où $\sum_{i=1}^{N_h} p_{hi} = 1$. Donc, nous sélectionnons en tout $n = \sum_{h=1}^L n_h$ UPF. Dans l'UPF (h, i) sélectionnée, nous sélectionnons n_{hi} unités secondaires d'échantillonnage (USE) avec remise et avec probabilité de sélection par tirage p_{hij} , où $\sum_{j=1}^{N_{hi}} p_{hij} = 1$ et N_{hi} est le nombre d'USE dans l'UPF (h, i) . Pour un item donné de l'enquête, soit X_h le total de population pour la strate h , et soit $X = \sum_{h=1}^L X_h$ le total global de population. Le total X peut correspondre à un total pour la population complète ou pour une sous-population particulière.

Notre objectif est de construire un intervalle de confiance pour le total X . Soit X_{hij} un estimateur sans biais de X_{hij} , le total de population pour l'unité secondaire j dans l'unité primaire i dans la strate h . Alors, un estimateur fondé sur le plan habituel de X est $\bar{Y} = \sum_{h=1}^L \bar{Y}_h$, où $\bar{Y}_h = \frac{1}{n_h} \sum_{i=1}^{N_h} p_{hi}^{-1} \bar{Y}_{hi}$, \bar{Y}_{hi} est un estimateur sans biais par rapport au plan de X_h fondé sur des données obtenues après de l'UPF i dans la strate h et $\bar{Y}_{hi} = \frac{1}{n_{hi}} \sum_{j=1}^{N_{hi}} p_{hij}^{-1} X_{hij}$ est un estimateur sans biais de X_{hi} dans la strate h .

Sous la condition standard voulant que l'échantillonnage soit indépendant d'une strate à l'autre, la variance de \bar{Y} peut s'écrire $V(\bar{Y}) = \sum_{h=1}^L V_h$, où $V_h = \text{Var}(\bar{Y}_h)$. Dans la suite de l'article, nous donnerons aux termes V_h le nom de variances au niveau de la strate et nous supposons que $n_h \geq 2$ pour tout $h = 1, 2, \dots, L$. Notons que V_h dépend du plan d'échantillonnage utilisé dans la strate h , et qu'il est distinct de la variance intra-strate des valeurs de X au niveau de l'élément. Un estimateur sans biais simple de $V(\bar{Y})$ est $\hat{V}(\bar{Y}) = \sum_{h=1}^L \hat{V}_h$, où $\hat{V}_h = n_h^{-1}(n_h - 1)^{-1} \sum_{i=1}^{N_h} (p_{hi}^{-1} \bar{Y}_{hi} - \bar{Y}_h)^2$; voir, par exemple, Wolter (1985, page 44). Soulignons que l'estimateur \hat{V}_h est un multiple d'une somme de différences quadratiques parmi les termes $p_{hi}^{-1} \bar{Y}_{hi}$ aléatoires $p_{hi}^{-1} \bar{Y}_{hi}$ suivront approximativement une loi En outre, sous des conditions de régularité, les variables aléatoires $p_{hi}^{-1} \bar{Y}_{hi}$ suivront approximativement une loi

solution de l'équation $2V(Y) - V(Y)^2 - V(Y)^2 = 0$ (1.1)

$$d = \sum_{h=1}^L \left(\sum_{i=1}^{N_h} (n_h - 1)^{-1} \bar{Y}_{hi}^2 \right) \{V(Y)\}^2 \quad (1.2)$$

où $\{V(Y)\} = \sum_{h=1}^L 2(n_h - 1)^{-1} \bar{Y}_h^2$ de \bar{Y}_h et de $V(Y)$ à $V(Y)$ dans l'expression (1.2) mène à l'estimateur de type Satterthwaite (1946) du nombre de degrés de liberté $d_s = \sum_{h=1}^L \left\{ \sum_{i=1}^{N_h} (n_h - 1)^{-1} \bar{Y}_{hi}^2 \right\} \{V(Y)\}^2$ (1.3)

Pour certains renseignements généraux sur d_s et les estimateurs connexes, consulter, par exemple, Smith (1936), Satterthwaite (1941, 1946), Cochran (1977, page 96) et Kendall, Stuart et Ord (1983, pages 91-92). Afin de construire des intervalles de confiance pour un paramètre de sous-population, Casady, Dorfman et Wang (1998) utilisent des notions bayésiennes pour élaborer des mesures connexes du nombre de degrés de liberté pour une statistique t de Student.

Sous les plans de sondage dans lesquels n_h est grand pour tout h , l'erreur dans l'estimation de V_h est relativement faible et d_s peut fournir un estimateur satisfaisant de l'expression (1.2). Cependant, dans de nombreuses enquêtes à grande échelle, n_h est petit, par exemple $n_h = 2$. Quand n_h est petit, la condition (C.1) et les opérations algébriques ordinaires donnent le résultat d'espérance $E(V_h^2) = (n_h - 1)^{-1}(n_h + 1)V_h^2$. Cela implique que l'estimateur de Satterthwaite classique du nombre de degrés de liberté d_s peut sous-estimer gravement d et que l'intervalle de confiance correspondant $\bar{Y} \pm t_{d_s, 1-\alpha/2} \{\hat{V}(\bar{Y})\}^{1/2}$ peut avoir un taux de couverture réel considérablement plus faible que le taux nominal $1 - \alpha$. Par conséquent, Jang (1996) a considéré un estimateur alternatif du nombre de degrés de liberté donné par $d_{ns} = (3L + 14)^{-1}(9L) \hat{d}_s$ (1.4)

pour le plan de sondage à deux UPF par strate.

Résumé

Les présents travaux ont été motivés par une étude de l'interférence concernant des sous-populations géographiques-concentrées dans la troisième National Health and Nutrition Examination Survey (NHANES III) réalisée aux Etats-Unis. Pour des renseignements généraux sur la NHANES III, consulter National Center for Health Statistics (1996). Dans de nombreuses analyses, les données de la NHANES III sont traitées comme émanant d'un plan d'échantillonnage à plusieurs degrés stratifié comportant 49 strates et 2 unités primaires d'échantillonnage (UPF) par strate. Par conséquent, les inférences formelles faites d'après les données de la NHANES III (par exemple, construction d'intervalles de confiance) reposent souvent sur l'hypothèse que les estimateurs de variance comme ceux sur basés sur environ 49 degrés de liberté et sont donc relativement stables.

Bibliographie

Australian Bureau of Statistics (ABS) (2008). *Employee Earnings and Hours*. Numéro de catalogue 6306.0.

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, 12, 470-482.

Chipperfield, J., et Preston, J. (2007). Bootstrap efficace pour les enquêtes-entreprises. *Techniques d'enquête*, 33, 187-193.

Estevao, V., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Fumakoka, F., Saigo, H., Sitter, R.R. et Toida, T. (2006). Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*, 32, 169-175.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, 181-184.

Kovar, J.G., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25-45.

McCarthy, P.J., et Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics (Séries 2 No 95)*, U.S. Government Printing Office.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

Särndal, C.-E., Swenson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

White, V., et Hayman J. (2006). Smoking behaviours of Australian secondary students in 2005. National Drug Strategy Monograph Series No. 59. Canberra: Australian Government Department of Health and Ageing.

Yeo, D., Mantel, H. et Liu T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

Au moyen de résultats standard sur l'espérance et la variance en ce qui concerne l'échantillonnage bootstrap EASSR et de certaines opérations algébriques fastidieuses mais simples, les composantes de l'estimateur de variance bootstrap sont données ci-après. L'espérance conditionnelle de Y^* sachant s_3 est

$$E_{j^*}(Y^*) =$$

$$\left(1 - \lambda_{1h} + \lambda_{1h} \frac{m_{1h}^*}{m_{1h}} \delta_{1hi} \sum_{h_1=1}^h \sum_{h_2=1}^{h_1} w_{h_1} w_{h_2} \right) \left(1 - \lambda_{2h_1} + \lambda_{2h_1} \frac{m_{1h}^*}{m_{1h}} \delta_{1hi} \sqrt{\frac{m_{1h}^*}{m_{1h}}} \delta_{2hi} \frac{m_{2h_1}^*}{m_{2h_1}} \delta_{2hj} \right) Y_{ij}^*$$

et la variance conditionnelle de Y^* sachant s_3 est

$$\text{Var}_{j^*}(Y^*) = \sum_{h=1}^H \frac{N_{1h}^*}{N_{1h}} \sum_{h_1=1}^h \frac{m_{1h}^*}{m_{1h}} \sum_{h_2=1}^{h_1} \frac{N_{2h_1}^*}{N_{2h_1}} \delta_{2hj_1} \delta_{2hj_2} \left(1 - f_{3hj_1} \right) s_{3hj_1}^2.$$

Les espérances conditionnelles de $E_{j^*}(Y^*)$ et de $\text{Var}_{j^*}(Y^*)$ sachant s_2 sont

$$E_{2^*}(E_{j^*}(Y^*)) = \sum_{h_1=1}^h \sum_{h_2=1}^{h_1} w_{h_1} \left(1 - \lambda_{1h} + \lambda_{1h} \frac{m_{1h}^*}{m_{1h}} \delta_{1hi} \right) Y_{hi}^*$$

$$E_{2^*}(\text{Var}_{j^*}(Y^*)) =$$

$$\sum_{h=1}^H \frac{N_{1h}^*}{N_{1h}} \sum_{h_1=1}^h \frac{m_{1h}^*}{m_{1h}} \sum_{h_2=1}^{h_1} \frac{N_{3h_1}^*}{N_{3h_1}} \delta_{3hi_1} \left(1 - f_{3hj_1} \right) s_{3hj_1}^2$$

et la variance conditionnelle de $E_{j^*}(Y^*)$ sachant s_2 est

$$\text{Var}_{2^*}(E_{j^*}(Y^*)) = \sum_{h=1}^H \frac{N_{1h}^*}{N_{1h}} \sum_{h_1=1}^h \frac{m_{1h}^*}{m_{1h}} \delta_{1hi} \frac{N_{2h_1}^*}{N_{2h_1}} \left(1 - f_{2hi} \right) s_{2hi}^2.$$

Enfin, les espérances conditionnelles de $E_{2^*}(\text{Var}_{j^*}(Y^*))$ et de $\text{Var}_{2^*}(E_{j^*}(Y^*))$ sachant s_1 sont

$$E_{1^*}(E_{2^*}(\text{Var}_{j^*}(Y^*))) =$$

$$\sum_{h=1}^H \frac{N_{1h}^*}{N_{1h}} \sum_{h_1=1}^h \frac{m_{1h}^*}{m_{1h}} \sum_{h_2=1}^{h_1} \frac{N_{2h_1}^*}{N_{2h_1}} \sum_{h_3=1}^{h_2} \frac{m_{3h_1}^*}{m_{3h_1}} \left(1 - f_{3hj_1} \right) s_{3hj_1}^2$$

$$E_{1^*}(\text{Var}_{2^*}(\text{Var}_{j^*}(Y^*))) = \sum_{h=1}^H \frac{N_{1h}^*}{N_{1h}} \sum_{h_1=1}^h \frac{m_{1h}^*}{m_{1h}} \sum_{h_2=1}^{h_1} \frac{N_{2h_1}^*}{N_{2h_1}} \left(1 - f_{2hi} \right) s_{2hi}^2$$

qui sont égales aux troisième et deuxième termes de (2.1) respectivement, et la variance conditionnelle de $E_{2^*}(E_{j^*}(Y^*))$ sachant s_1 est

$$\text{Var}_{1^*}(E_{2^*}(E_{j^*}(Y^*))) = \sum_{h=1}^H \frac{N_{1h}^*}{N_{1h}} \left(1 - f_{1h} \right) s_{1h}^2$$

qui est égale au premier terme de (2.1).

L'estimateur de variance à trois degrés standard sans biais (2.1) dans le cas où $\hat{\theta}$ est l'estimateur linéaire, Y^* = $\sum_{h=1}^H \sum_{t=1}^{n_{th}} \sum_{j=1}^{m_{th}} \sum_{k=1}^{n_{thj}} w_{hijk}^* y_{hijk}^*$. L'estimateur de variance bootstrap de Y^* est donné par :

$$\text{Var}(Y^*) = \text{Var}_1(E_{2*}(E_{3*}(Y^*)))$$

$$+ E_{1*}(\text{Var}_{2*}(E_{3*}(Y^*))) + E_{1*}(E_{2*}(\text{Var}_{3*}(Y^*)))$$

À la présente annexe, nous montrons que l'estimateur de variance bootstrap rééchantilloné à plusieurs degrés pour l'échantillonnage stratifié à trois degrés se réduit à

Annexe

rééchantonné à plusieurs degrés donne de meilleurs résultats que le bootstrap rééchantonné à un degré et que le bootstrap bernoullien à plusieurs degrés pour les statistiques lisses, telles que les moyennes, les ratios et les coefficients de corrélation et de régression.

Tableau 3 Racine relative de l'erreur quadratique moyenne (%) des estimateurs de variance

Moyenne (μ_y)		Moyenne (μ_z)		Ratio (R_y)	
EBRP	EBRU	EBRP	EBRU	EBRP	EBRU
Pop. I	31,9	44,6	31,7	32,3	37,3
Pop. II	33,9	38,2	33,4	32,2	32,9
Pop. III	33,8	35,9	33,0	33,0	33,1
Pop. IV	35,3	37,4	34,2	32,8	32,8
Pop. V	32,0	31,9	34,3	33,0	33,1
Pop. VI	16,4	40,7	16,4	16,5	47,5
Pop. VII	16,1	47,4	16,1	16,4	49,0
Pop. VIII	16,3	40,3	16,7	16,3	48,8
Pop. IX	19,2	26,7	19,3	20,0	28,6
Pop. X	19,8	22,4	19,9	20,2	29,0
Coefficient de corrélation (ρ_{yz})		Coefficient de régression (β_{yz})		Médiane (M_y)	
EBRP	EBRU	EBRP	EBRU	EBRP	EBRU
Pop. I	47,8	36,6	37,2	88,7	80,1
Pop. II	48,4	66,6	37,4	93,4	91,0
Pop. III	35,9	38,4	37,5	80,4	80,3
Pop. IV	42,6	45,4	38,0	40,3	96,6
Pop. V	40,3	40,0	37,3	40,1	30,7
Pop. VI	21,6	48,4	47,8	17,0	55,3
Pop. VII	21,4	48,4	49,0	16,8	51,4
Pop. VIII	21,6	46,3	48,6	16,9	49,7
Pop. IX	21,5	29,4	29,9	21,6	42,7
Pop. X	22,7	27,8	20,6	21,9	38,9

Tableau 4 Biases relatif (%) et racine relative de l'erreur quadratique moyenne (%) de l'estimateur de variance bootstrap rééchantonné

μ_y		R_{yz}		ρ_{yz}		β_{yz}		M_y	
Pop. II	-0,42	-0,29	-1,51	-0,08	20,98	Pop. IV	0,40	18,28	12,24
Pop. VII	-0,22	-0,24	-0,03	-0,28	7,24	Pop. IX	0,62	12,24	43,8
Racine relative de l'erreur quadratique moyenne (%)		Biases relatif (%)		Racine relative de l'erreur quadratique moyenne (%)		Notre : L'erreur de simulation la plus importante sur les biais relatifs était inférieure à 0,6 %.			
Pop. II	32,6	32,4	48,4	37,3	97,8	Pop. IV	32,8	99,4	50,0
Pop. VII	16,2	16,1	21,5	16,9	50,0	Pop. IX	19,1	43,8	20,5

Tableau 2
Biases relatif (%) des estimateurs de variance

	Moyenne (μ_y)			Moyenne (μ_z)			Ratio (R_y)		
	EBRP	EBRU	EBB	EBRP	EBRU	EBB	EBRP	EBRU	EBB
Pop. I	-0.28	-6.73	27.10	0.42	-6.63	27.32	0.00	-9.07	36.22
Pop. II	-0.05	-2.21	11.83	0.59	-1.64	12.54	-0.43	-9.26	36.40
Pop. III	-0.79	-2.63	3.62	-0.93	-2.66	3.40	-0.17	-5.30	5.19
Pop. IV	-0.23	-0.52	3.60	-0.18	-0.46	3.61	0.53	-4.65	5.98
Pop. V	0.15	-1.60	4.55	0.15	-1.64	4.54	0.22	-4.85	6.41
Pop. VI	0.70	-39.18	-0.34	0.65	-39.36	-0.28	1.57	-46.40	1.30
Pop. VII	0.19	-46.19	-0.26	-0.06	-46.48	-0.57	-0.27	-48.19	-0.73
Pop. VIII	0.37	-38.62	-0.41	0.23	-39.36	-0.46	-0.26	-47.93	-0.62
Pop. IX	0.42	-20.85	-7.76	-0.51	-20.03	-8.41	0.13	-23.13	-8.87
Pop. X	-0.56	-12.35	-6.08	0.70	-10.87	-6.93	-0.72	-23.70	-9.51
Pop. I	-2.31	-10.23	32.17	-0.08	-9.05	36.41	19.04	-19.86	33.21
Pop. II	-1.51	-8.41	29.65	0.05	-8.74	36.41	19.29	2.42	40.85
Pop. III	0.36	-4.37	5.69	0.05	-5.12	5.42	7.50	4.28	9.72
Pop. IV	2.18	-0.60	7.17	0.28	-5.05	5.70	17.40	16.17	34.37
Pop. V	0.79	-2.71	5.95	0.26	-5.40	6.34	8.29	4.78	11.49
Pop. VI	0.32	-46.67	0.14	0.89	-46.59	0.69	13.57	-33.56	9.15
Pop. VII	-0.07	-46.78	-0.29	-0.21	-47.85	-0.60	14.68	-33.16	11.86
Pop. VIII	0.31	-44.25	-0.27	-0.09	-47.54	-0.55	2.09	-38.90	-0.64
Pop. IX	-0.93	-23.02	-9.30	-0.20	-23.48	-9.20	8.08	-17.23	-1.97
Pop. X	-0.82	-19.35	-8.24	-1.02	-23.89	-9.75	2.10	-13.84	-5.46

Nota : L'erreur de simulation la plus importante sur les biais relatifs était inférieure à 0,7 %.

Dans le cas de statistiques non lissées, telles que les médianes, l'EBRP et l'EBB ont tous deux tendance à surestimer les variances de population réelles, tandis que l'EBRU a tendance à les sous-estimer. En outre, la racine relative de l'erreur quadratique moyenne est jusqu'à trois fois plus élevée pour les médianes que pour les moyennes. L'EBRP donne de meilleurs résultats que l'EBB pour les populations artificielles I à V, dans lesquelles les fractions d'échantillonnage de premier degré sont petites ($f_h = 0,1$), tandis que l'EBB donne d'un peu meilleurs résultats que l'EBRP pour les populations artificielles VI à X, dans lesquelles les fractions d'échantillonnage de premier degré sont plus grandes ($f_h = 0,3$ ou $0,5$). Cette surestimation du bootstrap rééchantonné à plusieurs degrés pour les médianes concorde avec les résultats présentés dans les études de Kovar et coll. (1988) et de Rao et coll. (1992) pour le bootstrap rééchantonné à un degré. Il convient de souligner que le bootstrap rééchantonné original, introduit par Rao et Wu (1988), a été élaboré uniquement pour des statistiques lissées, telles que les moyennes, les ratios, ainsi que les coefficients de corrélation et de régression.

Nous avons examiné l'EBRP en utilisant les poids d'estimation calés, $w_{hij} = w_{1h1}w_{2h2}$, qui satisfont la contrainte de calage $\sum_{(hij) \in U} w_{1h1}w_{2h2}x_{2hij} = X_2$, où $X_2 = \sum_{(hij) \in U} x_{2hij}$ est le total de population pour la variable auxiliaire de deuxième degré. Le tableau 4 donne le biais

5. Conclusion

Le présent article décrit l'extension de la méthode du bootstrap rééchantonné à l'échantillonnage à plusieurs degrés où les unités sont sélectionnées par échantillonnage aléatoire simple sans remise à chaque degré. Sous la méthode du bootstrap rééchantonné à plusieurs degrés proposée, les échantillons bootstrap sont sélectionnés sans remise et des facteurs d'échelle sont appliqués aux poids de sondage. Cette méthode proposée est relativement simple à mettre en œuvre et requiert un nombre considérablement plus faible de générations de nombres aléatoires que la méthode du bootstrap bernoullien général à plusieurs degrés. La méthode proposée convient aussi pour une vaste gamme de méthodes de pondération, y compris le calage, et d'ajustements pour tenir compte de la non-réponse et de la couverture insuffisante de la population. En outre, les résultats de l'étude par simulation Monte Carlo indiquent que le bootstrap

relatif et la racine relative de l'erreur quadratique moyenne de l'EBRP avec utilisation des poids d'estimation calés pour les quatre populations artificielles II, IV, VII et IX. Les biais relatifs et les racines relatives de l'erreur quadratique moyenne de l'EBRP avec des poids d'estimation calés sont semblables à ceux obtenus en utilisant les poids d'estimation non calés.

Les valeurs des paramètres que nous avons maintenues constantes dans les dix populations étaient $\mu_{x1h} = 25 \times (h + 1)$, $\sigma_{x1h}^2 = 10$, $\sigma_{x2h}^2 = 100$, $\rho_{x1h, x2h} = 0,75$ et $\rho_{y1h, y2h} = 0,50$. Les valeurs des paramètres que nous avons fait varier dans les dix populations étaient f_{1h} , les fractions d'échantillonnage de premier degré, f_{2hi} , les fractions d'échantillonnage de deuxième degré, p_{bih} et p_{w2hi} . Les valeurs de ces paramètres sont présentées au tableau 1.

Tableau 1
Caractéristiques des populations simulées

	f_{1h}	f_{2hi}	p_b	p_w
Pop. I	0,1	0,75	0,75	0,75
Pop. II	0,1	0,1	0,25	0,75
Pop. III	0,1	0,5	0,75	0,75
Pop. IV	0,1	0,5	0,25	0,75
Pop. V	0,1	0,5	0,25	0,25
Pop. VI	0,5	0,1	0,75	0,75
Pop. VII	0,5	0,1	0,75	0,25
Pop. VIII	0,5	0,1	0,25	0,25
Pop. IX	0,3	0,3	0,75	0,25
Pop. X	0,3	0,3	0,25	0,25

Les paramètres d'intérêt dans l'étude par simulation étaient la moyenne de population, μ_y , le ratio de population, $R_{yz} = \mu_y / \mu_z$, le coefficient de corrélation de population, $\rho_{yz} = \sigma_{yz} / \sigma_y \sigma_z$, le coefficient de régression de population, $\beta_{yz} = \sigma_{yz} / \sigma_z^2$ et la médiane de population, M_y . Afin d'estimer ces paramètres d'intérêt en utilisant les estimateurs de variance bootstrap à plusieurs degrés, nous avons tiré un total de $S = 20\,000$ échantillons aléatoires à deux degrés indépendants sans remise dans chacune des dix populations artificielles. En outre, nous avons tiré un total général de $T = 100\,000$ échantillons aléatoires simples à deux degrés indépendants sans remise dans chacune des dix populations artificielles afin d'estimer les variances de population réelles pour les paramètres d'intérêt. Pour calculer les estimateurs de variance bootstrap à plusieurs degrés, nous avons utilisé $B = 100$ échantillons bootstrap.

Pour comparer l'exactitude des estimateurs de variance bootstrap à plusieurs degrés, nous nous sommes servis du biais relatif (BR) et de la racine relative de l'erreur

$$\text{RREQM} = \frac{1}{\sqrt{\frac{1}{S} \sum_{s=1}^S (\text{Var}(Y'_s) - \text{Var}(Y))}} \quad \text{BR} = \frac{1}{\left[\frac{1}{S} \sum_{s=1}^S (\text{Var}(Y'_s) - \text{Var}(Y)) \right]}$$

où $\text{Var}(Y) = T^{-1} \sum_{t=1}^T (Y'_t - Y)^2$ est la variance de population réelle estimée, et $\text{Var}(Y'_s)$ est l'estimateur de variance bootstrap à plusieurs degrés pour le s^{e} échantillon de simulation.

Nous avons comparé l'estimateur de variance bootstrap rééchantonné à plusieurs degrés (EBRP) à l'estimateur de variance bootstrap rééchantonné à un degré (EBRU) et à l'estimateur de variance bootstrap bernoullien général (EBB) proposé par Funaoka et coll. (2006), pour des échantillons bootstrap utilisant les poids d'estimation non calés, $w_{hi} = w_{1hi} w_{2hi}$. Les tableaux 2 et 3 donnent le biais relatif et la racine relative de l'erreur quadratique moyenne des estimateurs EBRP, EBRU et EBB en utilisant des poids d'estimation non calés pour les dix populations artificielles.

Dans le cas de fonctions linéaires, telles que les moyennes, et de fonctions non linéaires, telles que les ratios, les coefficients de corrélation et les coefficients de régression, l'EBRP donne de meilleurs résultats que l'EBRU et l'EBB en ce qui concerne le biais relatif et la racine relative de l'erreur quadratique moyenne. Alors que l'EBRP donne systématiquement de bons résultats dans les dix populations artificielles, l'EBRU ne le fait que pour les populations artificielles III, IV et V, dans lesquelles les fractions d'échantillonnage de premier degré sont faibles ($f_{1h} = 0,1$) et celles de deuxième degré, grandes ($f_{2hi} = 0,5$), et l'EBB et celles de deuxième degré, grandes ($f_{2hi} = 0,5$), et l'EBB ne donne de bons résultats que dans les populations artificielles VI, VII et VIII, dans lesquelles les fractions d'échantillonnage de premier degré sont grandes ($f_{1h} = 0,5$) et celles de deuxième degré, faibles ($f_{2hi} = 0,1$). Ces fractions d'échantillonnage sont semblables aux fractions utilisées dans l'étude par simulation présentée dans Funaoka et coll. (2006). Les divers niveaux de corrélation entre les unités de premier degré, et entre les unités de deuxième degré à l'intérieur des unités de premier degré, dont il a été tenu compte en faisant varier les paramètres p_b et p_w , ont peu d'effet sur la performance des estimateurs de variance.

Les variables auxiliaires peuvent être utilisées pour former les estimateurs par calage de premier et de deuxième degrés :

$$I^{CAL1} = \sum_{(hi) \in s_1} w_{1hi} y_{1hi}$$
$$I^{CAL2} = \sum_{(hij) \in s_2} w_{12hij} y_{2hij}$$

où les poids de calage de premier degré, w_{1hi} , et les poids de calage de premier et de deuxième degrés combinés, w_{12hij} , sont donnés par :

$$w_{1hi} = w_{1hi} \left(1 + \left(X_1 - \sum_{(hi) \in s_1} w_{1hi} x_{1hi} \right)^T \right)$$

$$w_{12hij} = w_{1hi} w_{2hij} \left(1 + \left(X_2 - \sum_{(hij) \in s_2} w_{1hi} w_{2hij} x_{2hij}^T \right)^T \right)$$
$$\left(\sum_{(hij) \in s_2} w_{1hi} w_{2hij} x_{2hij}^T x_{2hij} \right)^{-1} \left(x_{2hij}^T \right)$$

La méthode du bootstrap rééchantillonné à plusieurs degrés peut alors être modifiée facilement de manière semblable pour traiter ces estimateurs par calage en remplaçant l'étape (d) de la procédure de la façon suivante :

(d) Calculer les poids bootstrap calés de premier et de deuxième degrés de la même manière que les poids calés de premier et de deuxième degrés :

$$w_{1hi}^* = w_{1hi} \left(1 + \left(X_1 - \sum_{(hi) \in s_1} w_{1hi}^* x_{1hi}^T \right)^T \right)$$

$$w_{12hij}^* = w_{1hi}^* w_{2hij}^* \left(1 + \left(X_2 - \sum_{(hij) \in s_2} w_{1hi}^* w_{2hij}^* x_{2hij}^T \right)^T \right)$$
$$\left(\sum_{(hij) \in s_2} w_{1hi}^* w_{2hij}^* x_{2hij}^T x_{2hij} \right)^{-1} \left(x_{2hij}^T \right)$$

Les estimations bootstrap calées de premier et de deuxième degrés sont calculées de la façon suivante :

4. Étude par simulation

Cette procédure peut être adaptée facilement à n'importe quel type de calage et étendue à n'importe quel nombre de degrés. Cette modification du bootstrap rééchantillonné tient compte des ajustements apportés aux poids de sondage à cause du calage. Idéalement, tous les ajustements des poids de sondage, y compris ceux dus à la non-réponse et à la couverture insuffisante de la population devraient également être appliqués aux poids bootstrap.

$$I^{CAL1*} = \sum_{(hi) \in s_1} w_{1hi}^* y_{1hi}$$
$$I^{CAL2*} = \sum_{(hij) \in s_2} w_{12hij}^* y_{2hij}$$

Pour examiner la performance de l'estimateur de variance bootstrap rééchantillonné à plusieurs degrés, nous avons procédé à une étude par simulation Monte Carlo. La simulation était limitée à l'échantillonnage stratifié à deux degrés et était fondée sur dix populations artificielles, qui étaient chacune subdivisée en $H = 5$ strates, avec $N_{2hi} = 50$ unités de premier degré dans chaque strate et $N_{2hi} = 40$ unités de deuxième degré dans chaque unité de premier degré.

En premier lieu, nous avons produit la variable auxiliaire de premier degré x_{1hi} par chaque unité de premier degré i dans la strate h à partir de la loi normale $N(\mu_{1hi}, (1 - p_{xib}) \sigma_x^2 / p_{xib})$. En deuxième lieu, nous avons produit la variable auxiliaire de deuxième degré, x_{2hij} , et les variables cibles de deuxième degré, y_{2hij} , et z_{2hij} , pour chaque unité de deuxième degré j à l'intérieur de l'unité de premier degré i dans la strate h à partir de la loi normale multivariée $N_3(\mu_{2hi}, \Sigma_{2hi})$, où μ_{2hi} est le vecteur des moyennes :

$$\mu_{2hi} = \begin{bmatrix} \mu_{x2hi} \\ \mu_{y2hi} \\ \mu_{z2hi} \end{bmatrix}$$

avec $\mu_{x2hi} = \mu_{y2hi} = \mu_{z2hi} = x_{1hi}$, et Σ_{2hi} est la matrice de variance-covariance :

$$\Sigma_{2hi} = \begin{bmatrix} \sigma_{x2hi}^2 & \rho_{xy2hi} \sigma_{x2hi} \sigma_{y2hi} & \rho_{xz2hi} \sigma_{x2hi} \sigma_{z2hi} \\ \rho_{xy2hi} \sigma_{x2hi} \sigma_{y2hi} & \sigma_{y2hi}^2 & \rho_{yz2hi} \sigma_{y2hi} \sigma_{z2hi} \\ \rho_{xz2hi} \sigma_{x2hi} \sigma_{z2hi} & \rho_{yz2hi} \sigma_{y2hi} \sigma_{z2hi} & \sigma_{z2hi}^2 \end{bmatrix}$$

avec $\sigma_{x2hi}^2 = \sigma_{y2hi}^2 = \sigma_{z2hi}^2 = (1 - \rho_{w2hi}) \sigma_x^2 / \rho_{w2hi}$.

Pour simplifier l'exposé, nous présentons le cas de l'échantillonnage stratifié à trois degrés, mais la méthode proposée peut être étendue facilement à n'importe quel nombre de degrés. La procédure du bootstrap rééchantonné sans remise pour l'échantillonnage stratifié à trois degrés est la suivante :

a) Tirer un échantillon aléatoire simple de n_{1h}^* UPE sans remise parmi les n_{1h}^* UPE dans l'échantillon. Soit δ_{1hi}^* égal à 1 si l'UPE i dans la strate h est sélectionnée et 0 autrement. Calculer les poids bootstrap des UPE :

$$w_{1hi}^* = \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}^*}{n_{1h}} \delta_{1hi}^* \right)$$

où $\lambda_{1h} = \sqrt{n_{1h}^* (1 - f_{1h}) / (n_{1h} - n_{1h}^*)}$.

b) Dans chaque UPE figurant dans l'échantillon, tirer un échantillon aléatoire simple de n_{2hi}^* USE sans remise parmi les n_{2hi}^* USE qui figurent dans l'échantillon. Soit δ_{2hi}^* égal à 1 si l'USE j dans l'UPE i dans la strate h est sélectionnée et 0 autrement. Calculer les poids bootstrap conditionnels des USE :

$$w_{2hij}^* = \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}^*}{n_{1h}} \delta_{1hi}^* \right) \left(1 - \lambda_{2hi} + \lambda_{2hi} \frac{n_{2hi}^*}{n_{2hi}} \delta_{2hij}^* \right)$$

$$\text{où } \lambda_{2hi} = \sqrt{n_{2hi}^* f_{1h} (1 - f_{2hi}) / (n_{2hi} - n_{2hi}^*)}.$$

c) Dans chaque USE figurant dans l'échantillon, tirer un échantillon aléatoire simple de n_{3hij}^* UFE sans remise parmi les n_{3hij}^* UFE qui figurent dans l'échantillon. Soit δ_{3hijk}^* égal à 1 si l'UFE k dans l'USE j dans l'UPE i dans la strate h est sélectionnée et 0 autrement. Calculer les poids bootstrap conditionnels des UFE :

$$w_{3hijk}^* = \left(1 - \lambda_{1h} + \lambda_{1h} \frac{n_{1h}^*}{n_{1h}} \delta_{1hi}^* \right) \left(1 - \lambda_{2hi} + \lambda_{2hi} \frac{n_{2hi}^*}{n_{2hi}} \delta_{2hij}^* \right) \left(1 - \lambda_{3hij} + \lambda_{3hij} \frac{n_{3hij}^*}{n_{3hij}} \delta_{3hijk}^* \right)$$

où $\bar{\theta} = \sum_{B=1}^B \bar{\theta}^{(B)} / B$.

Nous montrons en annexe que l'estimateur de variance bootstrap rééchantonné à plusieurs degrés pour l'échantillonnage stratifié à trois degrés défini par (3.1) se réduit à l'estimateur de variance à trois degrés sans biais standard (2.1) dans le cas où $\bar{\theta}$ est un estimateur linéaire. Le choix de $n_{1h}^* = \lfloor n_{1h} / 2 \rfloor$, $n_{2hi}^* = \lfloor n_{2hi} / 2 \rfloor$ et $n_{3hij}^* = \lfloor n_{3hij} / 2 \rfloor$ sera optimal et aura la propriété désirable de donner lieu à des poids bootstrap qui ne sont jamais négatifs.

La procédure proposée peut facilement être étendue à n'importe quel nombre de degrés en ajoutant des termes de la forme $-\lambda_{rh} (\prod_{i=1}^{r-1} \sqrt{n_i/n_i^*}) \delta_r^* + \lambda_{rh} (\prod_{i=1}^{r-1} \sqrt{n_i/n_i^*}) \delta_r^*$ à chaque degré, R , aux ajustements des poids bootstrap, où $\lambda_r = \sqrt{n_r^* (\prod_{i=1}^{r-1} f_i^*) (1 - f_r^*) / (n_r - n_r^*)}$.

Yeo, Mantel et Liu (1999) ont présenté une amélioration du bootstrap rééchantonné qui tient compte des ajustements apportés aux poids de sondage, tels que la poststratification. Par exemple, considérons un cas simple de calage non intégré au moyen d'information auxiliaire pour l'échantillonnage stratifié à deux degrés (Estevao et Samdal 2006) dont le double objectif est de produire des estimations pour une variable d'intérêt de premier degré, $Y_1 = \sum_{(hi) \in U} y_{1hi}$, ainsi qu'une variable d'intérêt de deuxième degré, $Y_2 = \sum_{(hij) \in U} y_{2hij}$. Supposons qu'il existe :

i) un ensemble de p variables auxiliaires de premier degré x_{1hi} pour lesquelles les totaux de population sont $\sum_{(hi) \in U} x_{1hi}$ sont connus, et où les totaux de population sont produits à partir d'une liste d'UPE pour lesquelles les x_{1hi} sont connus, où les totaux de population sont $\sum_{(hij) \in U} x_{2hij}$ sont connus, où les totaux de population sont obtenus auprès d'une source extérieure.

ii) un ensemble de q variables auxiliaires de deuxième degré x_{2hij} pour lesquelles les totaux de population sont connus pour chaque UPE dans la population ; et

Un estimateur sans biais de $\text{Var}(Y)$ est donné par (Sämdal, Swensson et Wretman 1992) :

$$\begin{aligned} \text{Var}(Y) = & \sum_{h=1}^{H-1} \frac{m_{1h}}{N_{1h}^2} (1 - f_{1h}) s_{1h}^2 \\ & + \sum_{h=1}^H \frac{N_{1h}}{N_{2h}} \sum_{k=1}^{m_{2h}} \frac{m_{1h}}{N_{2h}} (1 - f_{2hk}) s_{2h}^2 \\ & + \sum_{h=1}^H \frac{N_{1h}}{N_{2h}} \sum_{k=1}^{m_{2h}} \sum_{j=1}^{m_{3hj}} \frac{m_{1h}}{N_{2hj}} (1 - f_{3hjk}) s_{3hj}^2 \quad (2.1) \end{aligned}$$

$$\begin{aligned} \text{ou } f_{1h} = & (m_{1h}/N_{1h}), f_{2hk} = (m_{2hk}/N_{2hk}), f_{3hjk} = (m_{3hjk}/N_{3hjk}), \\ \bar{X}_h = & \sum_{i=1}^{m_{1h}} \bar{X}_{hi}/m_{1h}, \bar{X}_{hk} = \sum_{j=1}^{m_{2h}} \bar{X}_{hjk}/m_{2hk}, \bar{X}_{hjk} = \sum_{k=1}^{m_{3hj}} \bar{X}_{hjk}/m_{3hj}, \\ s_{1h}^2 = & \sum_{i=1}^{m_{1h}} (\bar{X}_{hi} - \bar{X}_h)^2 / (m_{1h} - 1), s_{2hk}^2 = \sum_{j=1}^{m_{2h}} (\bar{X}_{hjk} - \bar{X}_{hk})^2 / (m_{2hk} - 1) \\ \text{et } s_{3hjk}^2 = & \sum_{k=1}^{m_{3hj}} (\bar{X}_{hjk} - \bar{X}_{hjk})^2 / (m_{3hj} - 1). \end{aligned}$$

3. Bootstrap rééchantonné pour l'échantillonnage stratifié à plusieurs degrés

Rao et Wu (1988) ont proposé un changement d'échelle de la méthode du bootstrap standard pour divers plans d'échantillonnage, dont l'échantillonnage stratifié. Comme les facteurs d'échelle ajustent les valeurs des données d'enquête, cette méthode ne s'applique qu'à des statistiques lisses. Rao, Wu et Yue (1992) ont présenté une modification de cette méthode du bootstrap rééchantonné dans laquelle les facteurs d'échelle sont appliqués aux poids de sondage plutôt qu'aux valeurs des données. Cette méthode d'échantillonnage originale, mais à l'avantage supplémentaire d'être applicable à des statistiques non lisses ainsi qu'à des statistiques lisses. Kovar, Rao et Wu (1988) ont montré que si l'on utilise une taille d'échantillon bootstrap de $m_h = m_h - 1$, l'estimateur bootstrap rééchantonné donne de bons résultats pour les statistiques lisses.

Bien que les échantillons bootstrap soient habituellement sélectionnés avec remise, Chhaperclaud et Preston (2007) ont modifié la méthode du bootstrap rééchantonné pour l'appliquer à la situation où les échantillons bootstrap sont sélectionnés sans remise. Sous cette méthode du bootstrap rééchantonné sans remise, nous pouvons montrer que le choix de $m_h^* = \lfloor n_h/2 \rfloor$ ou $m_h^* = \lceil n_h/2 \rceil$ est optimal, où les opérateurs $\lfloor x \rfloor$ et $\lceil x \rceil$ arrondissent l'argument x vers le bas et vers le haut, respectivement, à l'entier le plus proche. Le choix de $m_h^* = \lfloor n_h/2 \rfloor$ à la propriété désirable de donner lieu à des poids bootstrap qui ne sont jamais négatifs.

2. Échantillonnage stratifié à plusieurs degrés

Dans le présent article, nous proposons une extension de la méthode du bootstrap rééchantonné à l'échantillonnage stratifié à plusieurs degrés dans lequel les unités sont sélectionnées par échantillonnage aléatoire simple sans remise à chaque degré. À la section 2, nous présentons la notation pour l'échantillonnage stratifié à plusieurs degrés. À la section 3, nous décrivons l'extension de l'estimateur bootstrap rééchantonné à l'échantillonnage à plusieurs degrés. À la section 4, nous présentons les principaux résultats d'une étude par simulation. Enfin, à la section 5, nous tirons certaines conclusions.

Le bootstrap bernoullien général à l'avantage de permettre le traitement de toute combinaison de tailles d'échantillon, mais il requiert un beaucoup plus grand nombre de générations de nombres aléatoires que le bootstrap bernoullien abrégé.

Pour simplifier, nous présentons le cas de l'échantillonnage stratifié à trois degrés. Considérons une population U divisée en H strates non chevauchantes $U = \{U_1, \dots, U_H\}$, où U_h est constituée de N_{1h} unités primaires d'échantillonnage (UPF). Au premier degré, un échantillon aléatoire simple sans remise (EASSR) de m_{1h} UPF est tiré avec probabilités de sélection $\pi_{1h} = N_{1h}/N_h$ dans chaque strate h . Supposons que l'échantillonnage d'échantillonnage (USE). Au deuxième degré, un EASSR de m_{2hi}/N_{2hi} dans chaque UPF sélectionnée. Supposons que l'USE j sélectionnée dans l'UPF i sélectionnée dans la strate h est constituée de N_{3hij} unités finales d'échantillonnage (UPF). Au troisième degré, un EASSR de m_{3hijk}/N_{3hijk} dans chaque UPF sélectionnée.

L'objectif est d'estimer le total de population $Y = \sum_{h=1}^H \sum_{i=1}^{m_{1h}} \sum_{j=1}^{m_{2hi}} \sum_{k=1}^{m_{3hij}} Y_{hijk}$ où Y_{hijk} est la valeur de la variable d'intérêt y pour l'UPF k dans l'USE j dans l'UPF i dans la strate h . Une estimation sans biais de Y est donnée par :

$$\hat{Y} = \sum_{h=1}^H \sum_{i=1}^{m_{1h}} \sum_{j=1}^{m_{2hi}} \sum_{k=1}^{m_{3hij}} Y_{hijk} = \sum_{h=1}^H \frac{N_{1h}}{m_{1h}} \sum_{i=1}^{m_{2hi}} \frac{N_{2hi}}{m_{2hi}} \sum_{j=1}^{m_{3hij}} Y_{hijk}$$

où $\bar{Y}_h = (N_{1h}/m_{1h}) \sum_{i=1}^{m_{2hi}} \bar{Y}_{hi}$, $\bar{Y}_{hi} = (N_{2hi}/m_{2hi}) \sum_{j=1}^{m_{3hij}} \bar{Y}_{hij}$ et $\bar{Y}_{hij} = (N_{3hij}/m_{3hij}) \sum_{k=1}^{m_{3hijk}} Y_{hijk}$. Cet estimateur peut également s'écrire sous la forme $\bar{Y} = \sum_{h=1}^H \sum_{i=1}^{m_{1h}} \sum_{j=1}^{m_{2hi}} \sum_{k=1}^{m_{3hijk}} W_{3hijk} Y_{hijk}$ où $W_{3hijk} = (N_{1h}/m_{1h}) (N_{2hi}/m_{2hi}) (N_{3hij}/m_{3hijk})$ est le poids d'échantillonnage pour l'UPF k dans l'USE j dans l'UPF i dans la strate h .

Bootstrap rééchantonné pour l'échantillonnage stratifié à plusieurs degrés

John Preston¹

Résumé

Dans les enquêtes par sondage de grande portée, il est fréquent d'employer des plans de sondage stratifiés à plusieurs degrés ou les unités sont sélectionnées par échantillonnage aléatoire simple sans remise à chaque degré. L'exécution de l'estimation de la variance sous ce genre de plan peut être assez fastidieuse, particulièrement pour les estimateurs non linéaires. Diverses méthodes bootstrap d'estimation de la variance ont été proposées, mais la plupart sont limitées à des plans à un seul degré ou à des plans en grappes à deux degrés. Nous proposons une extension de la méthode du bootstrap rééchantonné (Rao et Wu 1988) aux plans stratifiés à plusieurs degrés qui peut être adaptée facilement à n'importe quel nombre de degrés. Cette méthode convient pour une grande gamme de méthodes de reproduction, y compris la classe générale des estimateurs par cage. Nous avons réalisé une étude par simulation Monte Carlo pour examiner la performance de l'estimateur de variance bootstrap rééchantonné à plusieurs degrés.

Mots clés : Bootstrap ; cage ; échantillonnage à plusieurs degrés ; stratification ; estimation de variance.

1. Introduction

Les plans d'échantillonnage stratifié à plusieurs degrés conviennent particulièrement bien pour les enquêtes par sondage de grande portée, parce qu'ils ont l'avantage de permettre le regroupement des efforts de collecte. Diverses méthodes d'estimation de la variance existent pour ces plans d'enquête complexes. Les plus fréquentes sont la méthode de linéarisation (ou méthode de Taylor) et les méthodes de rééchantillonnage, telles que le jackknife, les répliques équilibrées répétées et le bootstrap. Dans le cas de plans de sondage complexes, l'exécution de la méthode de linéarisation peut être assez fastidieuse, car elle nécessite l'établissement de formules de variance distantes pour chaque estimateur non linéaire. Certaines approximations sont habituellement requises pour la variance de fonctions non linéaires, telles que les ratios, les coefficients de corrélation et les coefficients de régression, ainsi que les fonctionsnelles, telles que les quantiles.

En revanche, les diverses méthodes de rééchantillonnage s'appuient sur une seule formule de variance pour tous les estimateurs. Les méthodes de répétition de l'échantillon peuvent refléter les effets d'une vaste gamme de méthodes de reproduction, y compris le cage, et des ajustements dus à la non-réponse et à la couverture insuffisante de population. Les méthodes du jackknife et des répliques répétées équilibrées ne sont applicables qu'aux plans stratifiés à plusieurs degrés dans lesquels les grappes sont échantillonnées avec remise ou les fractions d'échantillonnage au premier degré sont négligables. Un certain nombre de méthodes bootstrap pour l'échantillonnage en population finie ont été proposées dans la littérature, y compris le bootstrap avec remise (McCarthy et Snowden 1985), le

bootstrap rééchantonné (Rao et Wu 1988), le bootstrap à concordance-miroir (*mirror-match bootstrap*) (Sitter 1992a) et le bootstrap sans remise (Gross 1980 ; Bickel et Freedman 1984 ; Sitter 1992b). Un résumé de ces méthodes bootstrap peut être consulté dans Shao et Tu (1995).

La plupart de ces méthodes bootstrap sont limitées à des plans à un seul degré ou à des plans à plusieurs degrés dans lesquels les unités d'échantillonnage de premier degré sont sélectionnées avec remise ou les fractions d'échantillonnage de premier degré sont faibles dans la plupart des strates. Cependant, dans de nombreuses enquêtes par sondage de grande portée, il est courant d'employer des plans de sondage à plusieurs degrés fortement stratifiés où les unités sont sélectionnées par échantillonnage aléatoire simple sans remise à chaque degré. Les enquêtes employeurs-employés, telles que la *Survey of Employee Earnings and Hours* (ABS 2008) et les enquêtes écoles-élèves, telles que la *National Survey on the Use of Tobacco by Australian Secondary School Students* (White et Hayman 2006), sont des exemples typiques de ce genre d'enquête.

McCarthy et Snowden (1985) ont proposé une extension de leur bootstrap avec remise à l'échantillonnage à deux degrés dans le cas particulier où les tailles de grappes sont égales et les tailles d'échantillon dans les grappes sont égales, tandis que Rao et Wu (1988) et Sitter (1992a) ont donné des extensions de leurs méthodes respectives du bootstrap rééchantonné et du bootstrap à concordance-miroir à l'échantillonnage en grappes à deux degrés. Récemment, Funakoka, Saigo, Sitter et Toida (2006) ont proposé des méthodes bootstrap de type bernoulli, le bootstrap bernoulli général et le bootstrap bernoulli abrégé, qui permettent de traiter facilement les plans stratifiés à plusieurs degrés dans lesquels les unités sont sélectionnées par

Remerciements

Le présent rapport est diffusé afin de tenir les parties

intéressées au courant des travaux de recherche et de favoriser les discussions. Les opinions exprimées concernant les questions statistiques et méthodologiques sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau. Les auteurs remercient deux examinateurs et un rédacteur associé de leurs commentaires et suggestions constructifs, ainsi que Carol Caldwell, Rita Petroni et Mark Sands de leurs commentaires utiles concernant une version antérieure du présent article. Les travaux de recherche de Jun Shao ont été financés en partie par la bourse SES-0705033 de la NSF.

Bibliographie

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429-440.
Hidiroglou, M.A., et Smdal, C.-E. (1998). Emploi des donnes *d'enqute*, 24, 11-20.
Kott, P. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.
Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
Oh, H.T., et Scheuren, F.J. (1983). Weighing adjustment of unit nonresponse. *Incomplete Data in Sample Surveys*. New York : Academic Press, 20, 143-184.
Smdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
Shao, J., et Stiel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.
Thompson, J.R. (2000). *Simulation: A Modeler's Approach*. New York : John Wiley & Sons, Inc.
Thompson, K.J., et Yung, W. (2006). To replicate (a weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.
Vartivarian, S., et Little, R.J. (2002). On the formation of weighing adjustment cells for unit non-response. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3553-3558.
Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.

Tableau 4
Résultats des simulations (en %) pour l'estimation de la variance sous mécanisme de réponse uniforme à l'intérieur de la strate

Estimation	Item	Branche d'activités			Ratio		
		V_{11}	V_{12}	V_{13}	V_{21}	V_{22}	V_{23}
Bâtiments	211000	BR -49,2	ST 55,4	ER 56,1	BR -50,4	ST 43,7	ER 43,9
	211000	BR -52,6	ST 58,0	ER 58,7	BR -53,7	ST 45,3	ER 45,4
	211000	BR -2,75	ST 29,8	ER 29,3	BR -19,2	ST 18,0	ER 18,4
	212300	BR -6,3	ST 30,3	ER 28,7	BR -8,96	ST 15,1	ER 14,9
	212300	BR -15,6	ST 15,3	ER 15,6	BR -12,0	ST 12,0	ER 12,0
	212300	BR -2,75	ST 29,8	ER 29,3	BR -19,2	ST 18,0	ER 18,4
	212300	BR -8,85	ST 32,0	ER 32,4	BR -9,03	ST 26,7	ER 26,8
	212300	BR -16,2	ST 16,0	ER 16,2	BR -14,7	ST 14,5	ER 14,8
	212300	BR -18,1	ST 18,0	ER 18,1	BR -18,2	ST 18,2	ER 18,2
	212300	BR -18,1	ST 18,0	ER 18,1	BR -18,2	ST 18,2	ER 18,2
	212300	BR -18,1	ST 18,0	ER 18,1	BR -18,2	ST 18,2	ER 18,2
	212300	BR -18,1	ST 18,0	ER 18,1	BR -18,2	ST 18,2	ER 18,2
Total	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
Bâtiments	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
Équipement	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
	211000	BR -8,12	ST 16,5	ER 16,7	BR -9,64	ST 12,4	ER 12,4
Total	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2
	211000	BR -27,8	ST 26,5	ER 26,5	BR -29,0	ST 24,2	ER 24,2

Tableau 3
Résultats des simulations (en %) pour l'estimation de la variance sous mécanisme de réponse dépendant de la covariable

Estimation	Item	Branche d'activités	v_1	v_2	v_3	v_4	v_5	v_6	v_7
Ratio	Total	211000	-35,8	-8,0	-10,3	47,4	48,6	252,9	113,9
Batiments	212300	ER	19,6	11,8	12,2	19,8	11,8	10,7	1,1
		BR	-20,4	-4,48	-22,2	-20,4	-2,69	54,8	266,4
		ST	30,3	26,8	31,1	26,8	27,3	139,1	6,3
		ER	12,6	9,9	12,6	12,6	9,6	6,3	0,1
	339900	BR	-21,2	0,26	-22,5	-21,2	1,55	-5,34	52,5
		ST	47,3	55,0	47,0	47,3	56,0	43,9	67,8
		ER	14,3	10,4	14,6	14,3	10,3	10,0	2,6
		BR	-20,7	3,83	-21,0	-20,7	4,08	-11,6	18,4
	541300	ST	32,7	34,9	32,8	32,7	35,0	29,4	32,0
		ER	12,6	10,4	12,3	12,6	10,3	6,9	0,1
		BR	-20,4	-6,31	-20,4	-20,4	-5,88	-10,9	39,8
		ST	42,5	42,3	42,7	42,5	42,3	39,9	64,0
Equipement	211000	ER	15,9	12,6	16,0	15,9	12,6	13,2	5,4
		BR	-20,1	0,09	-20,1	-20,1	0,33	-15,9	42,7
		ST	42,6	50,5	42,5	42,6	50,7	41,1	15,7
		ER	13,1	8,9	13,2	13,1	9,9	12,0	6,5
	212300	BR	-15,0	14,1	-17,3	-15,0	16,4	-9,37	27,9
		ST	63,9	87,7	62,6	63,9	90,0	64,1	69,6
		ER	16,2	13,3	16,7	16,2	13,0	14,7	6,7
		BR	-21,4	-4,13	-23,3	-21,4	-2,21	39,7	201,1
	339900	ST	31,7	28,4	32,5	31,7	29,0	113,7	204,4
		ER	13,3	10,2	13,5	13,3	10,1	7,7	0,2
		BR	-21,4	1,18	-22,8	-21,4	2,57	-7,29	50,8
		ST	51,2	60,8	50,9	51,2	61,9	47,9	69,2
Total	211000	ER	15,5	11,6	15,8	15,5	11,4	11,0	2,3
		BR	-19,7	6,16	-19,9	-19,7	6,43	-11,9	12,8
		ST	33,8	38,4	33,9	33,8	38,5	31,0	30,9
		ER	12,5	8,9	12,5	12,5	8,9	11,0	7,0
	212300	BR	-30,1	0,05	-31,9	-30,1	1,85	1,05	103,1
		ST	50,4	55,9	50,5	50,4	57,3	46,7	113,4
		ER	15,3	9,0	15,6	15,3	8,8	8,7	1,0
		BR	-33,2	-6,30	-34,6	-33,2	-4,96	17,6	204,5
	339900	ST	47,5	55,2	47,4	47,5	55,7	43,4	62,4
		ER	13,4	9,1	13,5	13,4	9,1	10,7	2,5
		BR	-22,9	1,68	-23,2	-22,9	1,94	-18,8	15,4
		ST	32,9	32,0	33,0	32,9	32,2	30,2	28,9
Batiments	211000	ER	11,5	7,2	11,6	11,5	7,1	10,6	5,2
		BR	-30,3	-0,15	-32,2	-30,3	1,65	-1,27	101,5
		ST	51,3	57,3	51,3	51,3	58,7	46,7	112,3
		ER	15,8	9,6	16,3	15,8	9,4	9,2	0,8
	212300	BR	-37,4	-13,5	-38,0	-37,4	-12,9	3,53	250,2
		ST	41,6	28,9	42,0	41,6	28,8	32,2	254,8
		ER	15,4	9,6	15,6	15,4	9,5	8,1	0,4
		BR	-20,3	-4,33	-20,3	-20,3	-4,00	-14,5	38,6
	339900	ST	42,3	42,4	42,4	42,3	42,4	40,1	62,8
		ER	14,6	11,9	14,7	14,6	11,8	13,6	5,0
		BR	-20,9	-0,54	-21,2	-20,9	-0,32	-18,9	14,5
		ST	41,6	47,8	41,6	41,6	48,0	40,6	40,9
Equipement	211000	ER	12,5	9,1	12,5	12,5	9,1	12,1	6,0
		BR	-17,8	11,2	-20,0	-17,8	13,3	-13,0	26,6
		ST	58,9	76,4	58,0	58,9	78,4	57,7	64,1
		ER	15,7	12,3	15,8	15,7	14,5	6,1	6,1
	212300	BR	-30,7	-3,27	-32,2	-30,7	-3,27	12,1	164,3
		ST	37,6	29,1	38,6	37,6	29,3	38,7	168,9
		ER	14,1	9,6	14,5	14,1	9,5	7,9	0,6
		BR	-24,1	2,52	-24,9	-24,1	3,27	-15,2	45,0
	339900	ST	51,2	61,0	51,1	51,2	61,5	47,7	64,1
		ER	14,8	9,9	15,1	14,8	9,8	11,9	2,3
		BR	-21,6	4,10	-21,9	-21,6	4,39	-18,1	10,1
		ST	33,6	35,2	33,7	33,6	35,3	31,5	28,2
Equipement	211000	ER	11,1	7,2	11,1	11,1	7,1	10,3	5,9
		BR	-17,8	11,2	-20,0	-17,8	13,3	-13,0	26,6
		ST	58,9	76,4	58,0	58,9	78,4	57,7	64,1
		ER	15,7	12,3	15,8	15,7	14,5	6,1	6,1
	212300	BR	-30,7	-3,27	-32,2	-30,7	-3,27	12,1	164,3
		ST	37,6	29,1	38,6	37,6	29,3	38,7	168,9
		ER	14,1	9,6	14,5	14,1	9,5	7,9	0,6
		BR	-24,1	2,52	-24,9	-24,1	3,27	-15,2	45,0
	339900	ST	51,2	61,0	51,1	51,2	61,5	47,7	64,1
		ER	14,8	9,9	15,1	14,8	9,8	11,9	2,3
		BR	-21,6	4,10	-21,9	-21,6	4,39	-18,1	10,1
		ST	33,6	35,2	33,7	33,6	35,3	31,5	28,2

Tableau 2
Caractéristiques de la population pour l'étude par simulation

Branches d'activités	Strate	Taille de la population	Fraction d'échantillonnage	Total	Corrélation avec la strate
211000	10	26	1,00	0,65	0,53
	2A	128	0,77	0,68	0,66
	2B	372	0,11	0,57	0,51
	2C	1 800	0,02	-0,07	0,00
	2D	10 406	0,00	0,28	0,00
	2E	30	1,00	0,96	0,95
212300	10	30	1,00	0,96	0,95
	2A	108	0,37	0,85	0,74
	2B	414	0,07	0,03	0,76
	2C	1 310	0,03	0,42	0,13
	2D	4 762	0,01	0,44	-0,22
	2E	158	1,00	0,80	0,40
339900	10	158	1,00	0,80	0,40
	2A	498	0,26	0,40	0,04
	2B	2 048	0,05	0,20	0,24
	2C	6 310	0,02	0,19	0,48
	2D	25 288	0,00	0,37	0,67
	2E	160	1,00	0,60	0,56
541300	10	160	1,00	0,60	0,56
	2A	959	0,38	0,20	0,39
	2B	4 531	0,06	0,28	0,13
	2C	17 913	0,01	0,08	0,06
	2D	67 440	0,00	0,13	-0,01
	2E	26	1,00	0,65	0,53

5. Conclusion

En présence de non-réponse dans les strates à tirage

complet (ou dans les strates dont la fraction d'échantillon-

nage est grande), l'estimateur de variance jackknife et

l'estimateur de variance par linéarisation qui ne tiennent pas

compte des strates à tirage complet (ou des strates dont la

fraction d'échantillonnage est grande) ne sont pas accep-

tables à cause de leur biais négatif important. Nous devons

deux estimateurs de variance asymptotiquement sans biais

et convergents en ajoutant un terme supplémentaire qui tient

compte de la variabilité due à la non-réponse dans les strates

à tirage complet (ou dans les strates dont la fraction

d'échantillonnage est grande). Nous devons également un

estimateur jackknife modifié qui est convergent quand les

strates à tirage complet sont les seules strates qui contribuent

à la variance due à la non-réponse (c'est-à-dire quand

l'hypothèse (6) est vérifiée).

Nos résultats de simulation montrent que les trois

estimateurs de variance donnent de bons résultats quand les

tailles d'échantillon de strate sont grandes et des résultats

inconsistants autrement, et que l'estimateur de variance

jackknife ne tenant compte d'aucune des fractions d'échan-

tillonnage est très prudent.

Comparativement à la méthode de linéarisation, le

jackknife requiert plus de ressources informatiques, mais il

offre d'autres avantages. Il est facile à programmer, ne

nécessite qu'une seule recette pour divers problèmes et ne

requiert pas de calculs compliqués ou distincts pour divers

estimateurs. Notre estimateur de variance par linéarisation

donné en (4) est en fait obtenu par linéarisation de l'esti-

mateur jackknife donné en (3).

Les tableaux 3 et 4 donnent, respectivement, les résultats de simulation sous les deux mécanismes de réponse. Ces résultats se résument comme il suit.

1. Deux estimateurs de variance ne tenant pas compte de V_2 , V_1 et V_L , possèdent en général un grand biais relatif négatif. Le taux d'erreur des intervalles de confiance connexe est également grand.
2. Deux estimateurs de variance convergents, V_L et V_{LL} , produisent des résultats fort semblables et sont généralement nettement meilleurs que V_1 et V_L en ce qui concerne le biais relatif et le taux d'erreur des intervalles de confiance connexe.
3. L'estimateur de variance jackknife V_J donne de bons résultats pour les branches d'activités 339900 et 541300, mais peut présenter un grand biais relatif positif dans les branches d'activités 211000 et 212300. Nous pensons qu'il s'agit d'un effet de « petit échantillon », puisque V_J est justifié par la convergence asymptotique et que la taille de la strate à tirage complet dans les branches d'activités 211000 et 212300 est égale à 26 et 30, respectivement (tableau 2). Pour les deux autres branches d'activités, la taille de la strate à tirage complet est de 158 et 160, respectivement. En fait, la perfor-

mance de V_L et V_{LL} est généralement meilleure pour les branches d'activités 339900 et 541300.

4. Dans certains cas, V_J possède un biais relatif négatif de plus de 10 %, qui est dû au fait que la fraction d'échantillonnage de certaines strates à tirage partiel est grande, autrement dit que l'approximation (6) ne tient pas suffisamment.
5. L'estimateur de variance jackknife V_J dans lequel sont ignorées toutes les fractions d'échantillonnage présente un biais relatif positif très grand et est trop prudent.

Tableau 1
Estimations de la variance pour \hat{Y} avec ajustement par le quotient dans l'enquête ACE-1

Branche d'activités	Item	Année	\hat{Y}	$\frac{V_L}{V_I}$	$\frac{V_L}{V_H}$	$\frac{V_L}{V_J}$	$\frac{V_L}{V_I}$	$\frac{V_L}{V_H}$	$\frac{V_L}{V_J}$
211000	Total	2002	1,63E+7	4,63E+11	0,97	1,14	1,17	1,00	17,3
		2003	2,28E+7	6,87E+12	0,95	1,21	1,26	1,00	2,81
		2004	2,30E+7	2,45E+12	0,98	1,23	1,25	1,00	4,77
		2005	3,08E+7	4,29E+12	0,98	1,22	1,24	1,19	4,77
		2006	4,18E+7	6,29E+12	0,99	1,17	1,19	1,00	8,78
	Bâtiments	2002	1,13E+7	3,99E+11	0,97	1,14	1,17	1,00	15,3
		2003	1,86E+7	5,78E+12	0,94	1,22	1,27	1,00	2,78
		2004	1,70E+7	8,39E+11	0,99	1,42	1,43	1,00	11,3
		2005	2,64E+7	3,84E+12	0,98	1,22	1,24	1,16	4,64
		2006	3,55E+7	5,41E+12	0,99	1,19	1,21	1,00	8,76
	Équipement	2002	3,20E+6	6,14E+10	0,98	1,15	1,17	1,00	7,26
		2003	4,18E+6	8,39E+11	0,97	1,22	1,24	1,00	1,70
212300	Total	2002	1,56E+6	4,14E+10	0,81	1,06	1,24	1,20	3,19
		2003	1,33E+6	1,21E+10	0,94	1,18	1,24	1,36	5,43
		2004	2,01E+6	2,86E+10	0,96	1,60	1,65	2,20	6,04
		2005	1,96E+6	1,93E+10	0,98	1,12	1,14	2,30	6,04
		2006	2,28E+6	2,19E+10	0,96	1,26	1,30	3,22	11,7
	Bâtiments	2002	2,22E+5	4,36E+8	1,00	1,11	1,11	1,64	8,61
		2003	1,49E+5	2,27E+8	0,96	1,28	1,32	1,48	7,32
		2004	4,14E+5	1,03E+8	0,96	46,6	46,6	75,3	426
		2005	2,23E+5	9,33E+8	0,99	1,12	1,13	1,32	1,88
		2006	2,20E+5	1,88E+9	0,97	1,20	1,22	1,19	2,29
	Équipement	2002	1,33E+6	4,05E+10	0,81	1,06	1,25	1,15	2,86
		2003	1,18E+6	1,33E+10	0,94	1,20	1,26	1,32	5,07
339900	Total	2002	1,75E+6	1,94E+10	0,99	1,27	1,29	1,10	3,71
		2003	1,58E+6	2,99E+10	0,98	1,24	1,27	1,10	1,60
		2004	1,70E+6	1,00E+10	0,99	1,40	1,40	1,69	4,61
		2005	1,77E+6	2,55E+10	0,99	1,28	1,29	1,25	3,02
		2006	1,94E+6	5,51E+10	0,99	1,23	1,25	1,12	2,15
	Bâtiments	2002	2,99E+5	1,21E+9	0,99	1,24	1,24	1,09	3,55
		2003	1,93E+5	8,54E+8	0,99	1,27	1,28	1,09	1,75
		2004	2,10E+5	2,00E+8	0,99	1,86	1,87	2,08	5,89
		2005	2,56E+5	5,07E+8	0,99	1,80	1,81	1,97	9,61
		2006	5,97E+5	4,93E+10	0,99	1,19	1,20	1,01	1,16
	Équipement	2002	1,45E+6	1,62E+10	0,99	1,27	1,28	1,07	3,02
		2003	1,39E+6	2,71E+10	0,97	1,24	1,27	1,09	1,58
541300	Total	2002	3,38E+6	2,32E+10	0,99	1,47	1,48	1,67	5,02
		2003	3,09E+6	2,61E+10	0,99	1,26	1,27	1,05	1,62
		2004	3,97E+6	1,12E+11	1,00	1,23	1,23	1,03	1,37
		2005	4,94E+6	2,54E+11	1,00	1,20	1,20	1,04	1,71
		2006	4,96E+6	2,82E+10	1,00	1,40	1,40	1,75	8,36
	Bâtiments	2002	7,41E+5	6,32E+9	1,00	1,70	1,71	1,64	7,47
		2003	4,29E+5	3,32E+9	1,00	1,29	1,29	1,01	1,33
		2004	6,96E+5	4,38E+10	1,00	1,22	1,22	1,00	1,40
		2005	7,12E+5	9,00E+9	1,00	1,25	1,25	1,08	2,08
		2006	8,73E+5	3,44E+9	1,00	1,58	1,59	1,63	9,88
	Équipement	2002	2,96E+6	1,39E+10	0,99	1,37	1,38	1,54	3,95
		2003	2,66E+6	1,94E+10	0,99	1,25	1,26	1,05	1,59
		2004	3,27E+6	5,83E+10	1,00	1,22	1,23	1,04	1,29
		2005	4,23E+6	2,40E+11	1,00	1,19	1,20	1,03	1,59
		2006	4,09E+6	2,35E+10	1,00	1,27	1,28	1,49	5,47

appréciable sur l'estimation de la variance. L'estimateur jackknife v_j , qui est corrigé de l'effet des strates à tirage complet, se situe généralement entre v_{j1} et v_{jL} . Dans certains cas, v_j est égal à v_{j1} ou en est très proche, indiquant que la variabilité due à la non-réponse provient principalement des strates à tirage partiel dont la fraction d'échantillonnage est grande. L'estimation jackknife v_j , qui ne tient pas compte des fractions d'échantillonnage, est beaucoup plus grande que n'importe quelle autre estimation.

4. Résultats des simulations

Ici, nous présentons une étude en simulation effectuée en nous servant de données modélisées d'après les branches d'activités de l'ACE-I présentées à la section précédente. À la section 4.1, nous décrivons les conditions de simulation et à la section 4.2, nous présentons et résumons les résultats.

4.1 Conditions de simulation

Nous avons modélisé notre population en utilisant les données des répondants recueillies pendant le cycle de collecte de 2003 pour les trois principaux items de dépenses visés par l'enquête (total, bâtiments et équipement). Les données pour la variable auxiliaire (paie) étaient disponibles pour toutes les unités dans la base de sondage. Les données pour la population complète ont été produites au moyen de l'algorithme SIMDAT (Thompson 2000) en choisissant comme cellules de modélisation les strates d'échantillonnage et comme taille de population, la taille originale dans la base de sondage pour chaque cellule. Le tableau 2 donne les fractions d'échantillonnage et les coefficients de corrélation avec la paie pour les données modélisées dans chaque strate.

Dans la simulation, des échantillons aléatoires simples stratifiés ont été tirés de la population produite. Nous examinons les propriétés statistiques des estimateurs de variance décrits à la section 2 appliqués à des échantillons répétés sous les deux mécanismes de réponse qui suivent, appliqués aux données de l'échantillon :

1. le mécanisme de réponse dépendant de la covariable obtenu par application aléatoire des propensions à répondre modélisées d'après les données d'enquête en prenant la paie comme covariable, ce qui produit de très fortes probabilités de réponse pour les grandes unités et de très faibles probabilités pour les petites unités (strates à tirage partiel) ;
2. le mécanisme de réponse uniforme à l'intérieur de la strate obtenu en utilisant les taux de réponse observés durant l'enquête comme probabilité de réponse à l'intérieur de la strate.

En moyenne, les probabilités de réponse par strate individuelle dans la branche d'activités était de 0,85, 0,76, 0,77, 0,76 et 0,68 pour les strates 10, 2A, 2B, 2C et 2D, respectivement.

Nous avons tiré 5 000 échantillons de la population, calculé \bar{Y} donné en (1) à partir de chaque échantillon en présence de non-réponse et d'ajustement des poids, et calculé la moyenne et la variance empirique des 5 000 valeurs de \bar{Y} . Nous avons répété l'exercice pour chaque branche d'activités et chaque item, en appliquant deux méthodes d'ajustement, à savoir l'estimateur par le ratio et l'estimateur de comptage. Quand \bar{Y} est l'estimateur par le ratio en utilisant la paie comme variable auxiliaire, la valeur absolue du biais relatif empirique est inférieure à 1,4 % et plus petite que 1 % dans la plupart des cas. Pour l'estimateur de comptage sous le mécanisme de réponse uniforme à l'intérieur de la strate, la valeur absolue du biais relatif empirique est inférieure à 0,5 %. L'estimateur de comptage n'est pas approximativement sans biais en théorie sous le mécanisme de réponse dépendant de la covariable. Toutefois, dans la simulation, la valeur absolue de son biais relatif empirique est inférieure à 1 % dans la plupart des cas et à une valeur maximale de 2,7 %. Nous avons utilisé la variance empirique de 5 000 valeurs de \bar{Y} comme « valeur réelle » de la variance de \bar{Y} .

4.2 Résultats

Dans 2 000 des 5 000 échantillons, nous avons calculé les six estimations de variance différentes pour les trois items, quatre branches d'activités et deux méthodes d'ajustement des poids. Nous avons examiné les propriétés statistiques de chaque estimateur de variance sur les échantillons répétés en nous servant du biais relatif (BR) défini comme

$$-1, \frac{\text{moyenne de 2 000 estimations de variance}}{\text{variance réelle}} - 1, \sqrt{\frac{\text{erreur quadratique moyenne empirique de l'estimation de variance}}{\text{variance réelle}}}$$

et le taux d'erreur (ER) défini comme étant la proportion empirique des intervalles de confiance à 90 % approximatifs $(\bar{Y} \pm 1,645 \sqrt{\text{estimation de la variance}})$ obtenus pour 2 000 échantillons qui ne contiennent pas le total de population réel.

année, dans les deux cas selon un plan d'échantillonnage aléatoire simple stratifié sans remise. L'échantillon ACE-1 comprend environ 75 % de l'échantillon de l'ACES (à peu près 46 000 entreprises sélectionnées par année pour l'ACE-1 et 15 000 sélectionnées par année pour l'ACE-2). Dans le plan d'échantillonnage ACE-1, les unités sont stratifiées par catégorie de taille dans chaque branche d'activités figurant dans la base de sondage. Il existe cinq strates ACE-1 distinctes dans chaque branche d'activités, soit une strate à tirage complet (appelée strate 10) et quatre strates à tirage partiel définies selon la taille de l'entreprise dans la branche d'activités (désignées par 2A à 2D, classées de la plus grande à la plus petite dans la branche d'activités), avec environ 500 strates à tirage partiel dans le plan de sondage de chaque année. Les fractions d'échantillonnage dans les strates correspondant aux unités de grande taille dans la branche d'activités (2A) peuvent être assez élevées : la plupart des années, environ 55 % de l'échantillon compris dans les strates 2A est tiré à un taux variant de 0,5 à 1. Dans les strates correspondant aux trois autres catégories de taille dans la branche d'activités, les fractions d'échantillonnage sont habituellement inférieures à 0,20. Les poids de sondage varient de 1 à 1 000, selon la branche d'activités et la catégorie de taille. Le composante ACE-2 est nettement moins stratifiée, le nombre de strates selon la catégorie de taille utilisées chaque année variant d'un total de 6 à 8, et les fractions d'échantillonnage étant inférieure à 0,01 dans toutes les strates. Notre analyse empirique est limitée à la composante ACE-1 de l'enquête qui répond à toutes les conditions décrites à la section précédente.

Le programme de l'ACES publie des estimations des totaux et des variations d'une année à l'autre. Ces estimations sont publiées pour l'ensemble de l'échantillon et selon le code de branche d'activités indiqué par les unités répondantes (qui ne correspond pas nécessairement au code de branche d'activités qui figure dans la base de sondage). En cas de non-réponse, les variantes sont estimées en utilisant la jackknife avec suppression d'un groupe (Kott 2001). Afin de tenir compte de la non-réponse totale, nous utilisons pour la composante ACE-1 la méthode d'ajustement par le quotient présentée à la section 2 en nous servant de données administratives sur la paie comme variable auxiliaire x .

Les classes de pondération sont les strates du plan de sondage, à condition qu'il y ait au moins un répondant dans la classe. La fusion des classes est extrêmement rare et est ignorée dans la suite du présent exposé. Des renseignements plus détaillés concernant le plan de sondage, la méthodologie et les limites des données de l'ACES sont disponibles en ligne à <http://www.census.gov/csd/ace>.

Même si le plan de sondage ACE-1 est relativement typique d'une enquête-entreprise, les données recueillies ne

3.2 Comparaisons

Il n'est pas évident qu'il y ait un effet de la méthode d'ajustement de la pondération pour corriger la non-réponse totale sur les erreurs-types pour la composante ACE-1, nous avons calculé des estimations de variance en utilisant des données ACE-1 corrigées de la non-réponse totale à l'aide de l'estimateur par le ratio en prenant la paie comme variable auxiliaire, dans quatre branches d'activités pour chacune desquelles les taux d'échantillonnage étaient élevés dans les strates à tirage partiel de grandes entreprises (2A). Les branches d'activités choisies sont représentatives des secteurs représentés dans l'ACES. Ces branches d'activités et leurs codes du Système de classification des industries de l'Amérique du Nord (SCIAN) sont : Extraction de pétrole et de gaz (211100), Extraction de minerais non métalliques (212300), Autres activités diverses de fabrication (339900), et Architecture, génie et services connexes (541300). Dans les tableaux et discussions qui suivent, les branches d'activités sont désignées par leur code du SCIAN.

Comme prévu, l'estimateur jackknife v_{j1} et l'estimateur jackknife par linéarisation v_{L1} sont très proches pour toutes les variables. Les estimateurs de variance convergents (v_{L1} et v_{j1}) produisent tous deux une valeur appréciablement plus grande que leurs analogues jackknife respectifs (v_{L1} et v_{j1}). En général, la plupart des dépenses en immobilisations sont déclarées par les entreprises des strates à tirage complet et des strates des grandes entreprises à tirage partiel, de sorte que l'introduction de la composante des non-répondants dans l'estimateur de variance a un effet

et $\bar{Y}^{(h)}$ l'analogue jackknife de \bar{Y} après la suppression de l'unité j dans $h \in \mathcal{C}$, quand nous traitons $X^{cp}K/X^{cp}$ comme des estimateurs. Alors, un estimateur jackknife de V_2 est donné par

$$V_{J2} = \sum_{h \in \mathcal{C}} \frac{N_h}{N_h - 1} \sum_{j \in \mathcal{F}_h} \left(\bar{Y}^{(hj)} - \frac{1}{N_h} \sum_{k \in \mathcal{F}_h} \bar{Y}^{(hk)} \right)^2$$

($n_h = N_h$ et $s_h = \mathcal{F}_h$ quand $h \in \mathcal{C}$). Dans la formule de \bar{Y} , le facteur $\sqrt{1 - X^{cp}/X^{cp}}$ fait la correction appropriée pour la non-réponse. Sous l'hypothèse P , $X^{cp}/X^{cp} \approx \pi^p$ est le taux de réponse, qui peut être considéré comme une fraction d'« échantillonnage » pour les strates à tirage complet. L'estimateur jackknife résultant de la variance totale $V_1 + V_2$ est alors $v_{J1} + v_{J2}$. Puisque $n_h = N_h$ (c'est-à-dire $1 - n_h/N_h = 0$) si la strate h est à tirage complet, il est facile de voir que $v_{J1} + v_{J2}$ est égal à

$$v_J = \sum_h \frac{n_h}{n_h - 1} \sum_{j \in \mathcal{F}_h} \left(\bar{Y}^{(hj)} - \frac{1}{n_h} \sum_{k \in \mathcal{F}_h} \bar{Y}^{(hk)} \right)^2, \quad (7)$$

où

$$\bar{Y}^{(hj)} = \begin{cases} \bar{Y}^{(hj)} & \text{si la strate } h \text{ est à tirage complet} \\ \sqrt{1 - \frac{n_h}{N_h}} \bar{Y}^{(hj)} & \text{si la strate } h \text{ n'est pas à tirage complet.} \end{cases}$$

Comparativement à l'estimateur de variance jackknife v_{J1} donné en (3), v_J donné en (7) tient compte de la variabilité due à la non-réponse dans les strates à tirage complet, tandis que v_{J1} ne le fait pas. Sous (6) et l'hypothèse M ou P , v_J est convergent. Enfin, l'estimateur jackknife où sont ignorées toutes les fractions d'échantillonnage est donné par :

$$v_J = \sum_h \frac{n_h}{n_h - 1} \sum_{j \in \mathcal{F}_h} \left(\bar{Y}^{(hj)} - \frac{1}{n_h} \sum_{k \in \mathcal{F}_h} \bar{Y}^{(hk)} \right)^2. \quad (8)$$

Cette estimateur semble être prudent, bien qu'il ne soit pas justifié théoriquement. Bref, nous avons les estimateurs de la variance totale $V_{ms}(\bar{Y})$ qui suivent :

1. L'estimateur jackknife v_{J1} défini en (3), qui produit une sous-estimation quand V_2/V_1 n'est pas négligéable ;
 2. L'estimateur par linéarisation v_{L1} défini en (4), qui ble ;
 3. $v_L = v_{L1} + v_{L2}$ avec v_{L2} défini en (5), qui est est asymptotiquement équivalent à v_{J1} ;
- convergent ;

À la présente section, nous appliquons les estimateurs de variance décrits à la section 2 à des données empiriques couvrant une période de cinq ans provenant de la composition de l'emploi de l'ACES présentée à la section 1. La section 3.1 donne le contexte des variables d'analyse, du plan d'échantillonnage et des méthodes d'estimation de l'ACES. La section 3.2 présente les comparaisons empiriques.

3.1 Contexte de l'ACES

L'ACES est conçue pour recueillir des données sur la nature et sur le niveau des dépenses en immobilisations des entreprises non agricoles exploitées aux États-Unis. Les répondants déclarent les dépenses en immobilisations, ventiles selon le type (dépenses en bâtiments et dépenses en équipement) pour l'année civile, dans toutes les succursales et divisions, pour toutes les opérations ayant lieu aux États-Unis. L'univers de l'ACES englobe deux sous-populations, à savoir les entreprises employées (ACES-1) et les entreprises non employées (ACE-2). (Une entreprise non employée est une entreprise qui ne possède pas d'employés rémunérés, dont les recettes d'affaires annuelles sont égales ou supérieures à 1 000 \$ (1 \$ ou plus dans le secteur de la construction) et qui est assujettie aux lois sur l'impôt fédérales. La plupart des non-employeurs sont des travailleurs autonomes exploitant de très petites entreprises non constituées en société qui peuvent ou non être la source principale de revenu du propriétaire.) Diverses questionnaires sont envoyés par la poste aux unités échantillonnées selon qu'il s'agit d'une entreprise ACE-1 ou ACE-2. De nouveaux échantillons ACE-1 et ACE-2 sont sélectionnés chaque

Sous échantillonnage aléatoire simple stratifié et l'hypothèse P , v_L est approximativement le même que l'estimateur de variance obtenu en traitant l'ensemble des répondants comme une phase supplémentaire de l'échantillonnage à deux phases) et en appliquant la formule de variance classique (quand $x_{hj} = 1$) ou la formule de variance totale des estimateurs par calage (Kott 1994, Sæmål, Swensson et Wretman 1992, et Hidiroglou et Sæmål 1998). Cet estimateur de variance n'est toutefois pas convergent quand l'hypothèse P n'est pas vérifiée.

4. $v_{JL} = v_{J1} + v_{J2}$, qui est asymptotiquement équivalente à v_L ;
5. L'estimateur de variance jackknife v_J défini en (7), qui est convergent quand l'expression (6) est vérifiée ;
6. L'estimateur jackknife v_{J1} .

$$V^m = X^{\frac{d}{d}} X^{\frac{d}{d}} \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \left(\right)$$

$$= E^m \left[V^m \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) \right] + V^m \left[E^m \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) \right]$$

$$= E^m \left[V^m \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) \right] = E^m \left[\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} V^m (Y^{\frac{d}{d}}) - 2 \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} V^m (Y^{\frac{d}{d}}) + V^m (Y^{\frac{d}{d}}) \right]$$

$$= E^m \left[\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \sigma^2 X^{\frac{d}{d}} - 2 \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \sigma^2 X^{\frac{d}{d}} + \sigma^2 X^{\frac{d}{d}} \right]$$

$$= \sigma^2 E^m \left[\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - X^{\frac{d}{d}} \right] = \sigma^2 E^m \left[\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - X^{\frac{d}{d}} \right]$$

Sous l'hypothèse P, soit V^m la variance par rapport aux I_{hj} .

Puisque

$$E^m \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) = 0,$$

nous obtenons

$$V^m \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) \approx E^m \left[V^m \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) \right]$$

$$\approx E^m \left[1 - \pi^{\frac{d}{d}} \sum_{h \in F_h} \sum_{j \in F_h} \delta_{hj} \left(Y_{hj}^{\frac{d}{d}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} X_{hj}^{\frac{d}{d}} \right) \right]$$

$$\approx E^m \left[\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - X^{\frac{d}{d}} \right] \left[S^2 \right]$$

où

$$S^2 = \frac{1}{\sum_{h \in F_h} \sum_{j \in F_h} \delta_{hj}} \left(Y_{hj}^{\frac{d}{d}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} X_{hj}^{\frac{d}{d}} \right)^2.$$

Comme $X^{\frac{d}{d}}$ et $X^{\frac{d}{d}}$ peuvent être estimés par $X^{\frac{d}{d}}$ et $X^{\frac{d}{d}}$ respectivement, pour estimer V^m nous devons uniquement trouver une estimateur de σ^2 ou de S^2 . Sous l'hypothèse M, un estimateur par la régression de $\beta^{\frac{d}{d}}$ est $Y^{\frac{d}{d}}/X^{\frac{d}{d}}$ et un estimateur convergent de σ^2 fondé sur les résidus de régression est

$$Y^{\frac{d}{d}} = \sum_{h \in F_h} \sum_{j \in F_h} \delta_{hj} \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right) \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right)$$

De même, sous l'hypothèse P, le résultat tient quand V^m est remplacé par V^m . Par conséquent, nous pouvons appliquer le jackknife à l'estimateur $X^{\frac{d}{d}} Y^{\frac{d}{d}} / X^{\frac{d}{d}}$. Soit

$$E^m = \sum_{h \in F_h} \sum_{j \in F_h} \delta_{hj} \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right) \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right).$$

$$V^m \approx E^m V^m - Y^{\frac{d}{d}} \left[\sum_{h \in F_h} \sum_{j \in F_h} \delta_{hj} \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right]$$

et C représente la série d'indices des strates à tirage complet. Un estimateur par le jackknife convergent de V^m peut être obtenu comme il suit. Notons que $X^{\frac{d}{d}}$, $X^{\frac{d}{d}}$ et $X^{\frac{d}{d}}$ sont des estimateurs, puisque $h \in F_h$ pour $h \in C$, mais que $X^{\frac{d}{d}}$ n'est pas un estimateur à cause de la non-réponse. Donc, nous ne pouvons pas appliquer le jackknife à la fonction $X^{\frac{d}{d}} X^{\frac{d}{d}} / X^{\frac{d}{d}} - Y^{\frac{d}{d}}$. Partant du calcul précédent, nous notons, sous l'hypothèse M, que

$$X^{\frac{d}{d}} = \sum_{h \in C} \sum_{j \in F_h} \delta_{hj} X_{hj}^{\frac{d}{d}} = \sum_{h \in C} \sum_{j \in F_h} \delta_{hj} I_{hj} X_{hj}^{\frac{d}{d}}$$

$$Y^{\frac{d}{d}} = \sum_{h \in C} \sum_{j \in F_h} \delta_{hj} Y_{hj}^{\frac{d}{d}} = \sum_{h \in C} \sum_{j \in F_h} \delta_{hj} I_{hj} Y_{hj}^{\frac{d}{d}}$$

où l'indice inférieur c désigne les strates à tirage complet,

$$(6) \quad V^m \approx V^m \left[\sum_{h \in C} \sum_{j \in F_h} \delta_{hj} \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right) \right],$$

et la linéarisation. Dans certaines applications, $\sum_{h \in C} \sum_{j \in F_h} N_h$ est négligeable et la non-réponse dans les strates à tirage partiel contribue de manière négligeable à la composante de variance V^m , c'est-à-dire

$$(5) \quad V^m \approx \sum_{h \in C} \sum_{j \in F_h} \delta_{hj} \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - Y^{\frac{d}{d}} \right)^2$$

donné par

Selon la théorie de l'échantillonnage, $\hat{\sigma}^2$ est aussi un estimateur convergent de S^2 sous l'hypothèse P. Donc, sous l'hypothèse M ou P, un estimateur convergent de V^m est

$$\hat{\sigma}^2 = \frac{1}{\sum_{h \in C} \sum_{j \in F_h} \delta_{hj}} \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right) \left(\frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} - \frac{X^{\frac{d}{d}}}{Y^{\frac{d}{d}}} \right)$$

donné par (2) quand les classes de pondération coïncident avec les strates. Comme V_{j1} , V_{L1} est convergent pour V_1 quand $k_p \rightarrow \infty$ sous l'hypothèse M ou P, ce qui découle du résultat classique pour le cas de données complètes (Krewski and Rao 1981). Puisque le ratio est une fonction lisse, sous l'hypothèse M ou P,

$$E_s(Y) = \sum_{j \in J_s} E_s(X_j^p) \approx \sum_{j \in J_s} \left(\frac{E_s(X_j^p) E_s(Y_j^p)}{E_s(X_j^p)} \right) = \sum_{j \in J_s} E_s(X_j^p),$$

où

$$\begin{aligned} X^p &= \sum_{j \in J_s} \delta^{p h_j} X^{h_j}, \\ X^{pr} &= \sum_{j \in J_s} \sum_{h \in J_h} \delta^{p h_j} I_{h_j} Y^{h_j}, \\ Y^{pr} &= \sum_{j \in J_s} \sum_{h \in J_h} \delta^{p h_j} I_{h_j} Y^{h_j}. \end{aligned}$$

et \mathcal{P}^h est la population finie dans la strate h . Posons que X_p est identique à X^p avec x_{h_j} remplacé par y_{h_j} . Alors

$$V_z = V^m[E_s(Y) - Y] \approx \sum_{j \in J_s} V^m \left(\frac{X^{pr}}{X^{p_r}} - Y^p \right).$$

Notons que V_z est petite si le taux de non-réponse est faible ($V_z = 0$ en l'absence de non-réponse) ou si le modèle sous l'hypothèse M est hautement prédictif. Si la fraction d'échantillonnage globale, $\sum_h n_h / \sum_h N_h$, converge vers 0, alors V_z / V_1 converge vers 0 et, par conséquent, V_{L1} et V_{j1} sont des estimateurs convergents de la variance totale $V_{ms}(Y)$. Notons que V_1 ne contient pas la variation due à la non-réponse émanant des strates à tirage complet. Comme, dans de nombreuses enquêtes, les valeurs de X provenant des strates à tirage complet sont influentes, dans le total Y et qu'il est difficile, dans les applications, de dire à quel point $\sum_h n_h / \sum_h N_h$ doit être petite pour qu'il ait lieu la convergence $V_z / V_1 \rightarrow 0$, il est nécessaire d'estimer V_z .

Sous l'hypothèse M, soit \bar{E}_m , V_m et \bar{C}_m l'espérance, la variance et la covariance conditionnelles, respectivement, sachant toutes les valeurs de x et tous les indicateurs de réponse. Puisque

$$\bar{E}_m \left(\frac{X^{pr}}{X^{p_r}} - Y^p \right) = 0,$$

nous obtenons

$$V_{j1} = \sum_{j \in J_s} \left(1 - \frac{n_h}{N_h} \right) \sum_{k \in J_h} \left(Y^{(h_j)} - \frac{1}{n_h} \sum_{k \in J_h} Y^{(h_k)} \right)^2 \quad (3)$$

l'analogie jackknife de Y quand l'unité j est supprimée dans la strate h . Notons que les fractions d'échantillonnage sont intégrées dans cette formule. Quand $k_p \rightarrow \infty$ pour toutes les classes de pondération, le résultat standard pour les cas de données complètes (voir, par exemple, Krewski et Rao 1981) implique que l'estimateur jackknife V_{j1} est convergent pour la variance d'échantillonnage $V_s(Y)$, sous l'hypothèse M ou P. Puisque V_1 est l'espérance de $V_s(Y)$, V_{j1} est également convergent pour V_1 sous certaines conditions mineures.

où X est le total de population finie des valeurs de y_j et la variance totale est donnée par

$$V_{ms}(Y) = E_m[V_s(Y)] + V[E_s(Y) - Y].$$

Puisque, dans (1), la fonction est la somme de ratios et que les données dans les différentes classes de pondération sont indépendantes, il est possible de dériver un estimateur par linéarisation de $V_s(Y)$ en utilisant le développement en série de Taylor. Quand les classes de pondération coïncident avec les strates, par exemple, Y est donné par (2) et est un estimateur par le ratio distinct dont l'estimateur de variance par linéarisation peut être obtenu en utilisant des méthodes standard. Un autre moyen d'établir un estimateur de variance par linéarisation consiste à linéariser l'estimateur V_{j1} (Thompson et Yung 2006). L'estimateur résultant est

$$V_{j1} = \sum_{j \in J_s} \sum_{h \in J_h} \left[\frac{X^{pr}}{X^{p_r}} (\bar{e}^{ph_j} - w_{h_j} e^{ph_j} I_{h_j} \delta^{ph_j}) + \frac{X^{pr}}{Y^p} (\bar{x}^{ph} - w_{h_j} x_{h_j} \delta^{ph_j}) \right]^2, \quad (4)$$

où $e^{ph_j} = Y^{ph_j} - (Y^{pr}/X^{pr}) x^{ph_j}$, $\bar{e}^{ph_j} = n^{-1} \sum_{j \in J_s} w_{h_j} e^{ph_j} I_{h_j} \delta^{ph_j}$ et $\bar{x}^{ph} = n^{-1} \sum_{j \in J_s} w_{h_j} x_{h_j} \delta^{ph_j}$. L'estimateur donné en (4) est exactement le même que l'estimateur de variance par linéarisation classique pour l'estimateur par le ratio distinct

exister une non-réponse et x_{hj} une covariable qui prend des valeurs positives et ne présente pas de non-réponse, j étant l'indice de l'unité de population et h , l'indice de strate. En suivant la trajectoire échantillon-réponse étudiée par Fay (1991), ainsi que par Shao et Steel (1999), nous considérons la population finie comme un recensement avec des valeurs y et x et des non-répondants, c'est-à-dire que chaque unité j dans la strate h de la population finie est associée à un indicateur $I_{hj} = 1$ si y_{hj} est un répondant et $= 0$ si y_{hj} est un non-répondant. Notre échantillon est tiré de cette population finie et, si l'unité j dans la strate h est comprise dans l'échantillon, y_{hj} est un répondant si $I_{hj} = 1$ et un non-répondant si $I_{hj} = 0$.

Soit E_j et V_j l'espérance et la variance, respectivement, sous l'échantillonnage, ainsi que E_m , V_m et F_m l'espérance, la variance et la probabilité, respectivement, sous le modèle m spécifié dans l'une des hypothèses suivantes.

Hypothèse M. Les valeurs de (y_{hj}, x_{hj}, I_{hj}) dans la population finie sont produites indépendamment à partir d'un modèle de superpopulation m . La population finie est divisée en P sous-populations telles que, dans la sous-population p , la probabilité de réponse est $F_m(I_{hj}, x_{hj}) = P_m(I_{hj} = 1 | x_{hj}) > 0$, $E_m(V_{hj} | x_{hj}) = \beta_p^m(y_{hj} | x_{hj})$ et $V_m(y_{hj} | x_{hj}) = \sigma_p^m(x_{hj})$, où β_p^m et σ_p^m sont des paramètres inconnus qui dépendent de p .

Hypothèse P. La population finie est divisée en P sous-populations telles que, sous un modèle de superpopulation, $F_m(I_{hj}) = 1 | y_{hj}, x_{hj}) = \pi^p > 0$ est constante dans la sous-population p .

La sous-population considérée dans l'hypothèse M ou dans l'hypothèse P est appelée classe de pondération pour la correction de la non-réponse (ou, brièvement, classe de pondération), puisque nous traitons les non-répondants par ajustement des poids dans chaque classe de pondération. (Si l'on procède à une imputation dans chaque sous-population, les sous-populations sont appelées classes d'imputation.) Dans les applications, les classes de pondération peuvent être des strates ou des unions de strates (les strates sont fusionnées quand elles contiennent un nombre insuffisant de répondants), où elles peuvent recouper diverses strates. L'hypothèse M comporte un modèle de prédiction reliant y_{hj} et x_{hj} , ainsi qu'un mécanisme de réponse dépendant d'une covariable à l'intérieur de chaque classe de pondération. Le mécanisme de réponse sous l'hypothèse P est le mécanisme de réponse uniforme à l'intérieur de la cellule de pondération, qui est souvent appelé modèle de réponse presque aléatoire. L'hypothèse P est plus forte que l'hypothèse M en ce qui a trait au mécanisme de réponse. Cependant, l'hypothèse M requiert un modèle explicite de la

relation entre y_{hj} et x_{hj} à l'intérieur de chaque classe de pondération. Ici, nous supposons que soit l'hypothèse M, soit l'hypothèse P est vérifiée. Les estimateurs qui peuvent être justifiés sous l'hypothèse P sont appelés estimateurs sous « quasi-randomisation » (Oh et Scheuren 1983). Quand nous étudions la convergence asymptotique des estimateurs, nous considérons le processus limite de $k_p \rightarrow \infty$ pour toute sous-population p avec H et P fixes, où k_p est la taille d'échantillon dans la classe de pondération p . Si les classes de pondération sont les mêmes que les strates ou les unions de strates, $k_p \rightarrow \infty$ équivaut à $n_h \rightarrow \infty$ pour toute strate h .

Après l'ajustement par le quotient pour corriger la non-réponse, nous considérons l'estimateur qui suit du total des valeurs de y dans la population finie :

$$\hat{Y} = \sum_p \sum_h \sum_{j \in s_h} \left(\frac{\hat{X}_p}{\hat{X}_p} w_{hj} \right) \delta_{pjh} I_{hj} y_{hj} = \sum_p \frac{\hat{X}_p}{\hat{X}_p} \hat{Y}_p, \quad (1)$$

où p est l'indice désignant la classe de pondération, s_h est l'échantillon dans la strate h , δ_{pjh} est l'indicateur pour la classe de pondération p , et w_{hj} est le poids de sondage construit pour l'échantillonnage stratifié.

$$\hat{X}_p = \sum_{j \in s_h} w_{hj} \delta_{pjh} x_{hj}, \quad \hat{X}_p = \sum_{j \in s_h} w_{hj} \delta_{pjh} I_{hj} x_{hj},$$

et

$$\hat{Y}_p = \sum_{j \in s_h} w_{hj} \delta_{pjh} I_{hj} y_{hj}.$$

Dans le cas particulier où les classes de pondération correspondent aux strates,

$$\hat{Y} = \sum_h \frac{\hat{X}_h}{\hat{X}_h} \hat{Y}_h, \quad (2)$$

où

$$\hat{X}_h = \sum_{j \in s_h} w_{hj} x_{hj}, \quad \hat{X}_h = \sum_{j \in s_h} w_{hj} x_{hj} I_{hj},$$

et

$$\hat{Y}_h = \sum_{j \in s_h} w_{hj} y_{hj} I_{hj}.$$

Quand la covariable $x_{hj} = 1$, \hat{Y} est appelée estimateur de comptage. L'estimateur de comptage contrôle les estimations des répondants par rapport aux totaux de population de la base de sondage. Quand les classes de pondération coïncident avec les strates, l'estimateur de comptage utilise les taux de réponse par classe non pondérés, comme le recommandent Varian et Little (2002).

Sous l'hypothèse M ou P,

Estimation de la variance en présence de non-répondants et de strates à tirage complet

Jun Shao et Katherine J. Thompson¹

Résumé

Les enquêtes-entreprises sont souvent réalisées selon un plan d'échantillonnage aléatoire simple stratifié à un degré sans remises comportant certaines strates à tirage complet. Bien que l'on recoure habituellement à l'ajustement de la pondération pour traiter la non-réponse totale, la variabilité due à la non-réponse est parfois omise en pratique quand on estime les variances. Cette situation pose surtout problème lorsque il existe des strates à tirage complet. Nous élaborons des estimateurs de variance qui sont convergents quand le nombre d'unités échantillonnées est grand dans chaque classe de pondération, en utilisant les méthodes du jackknife, de la linéarisation et du jackknife modifié. Nous commençons par appliquer les estimateurs ainsi obtenus à des données empiriques provenant de l'Annual Capital Expenditures Survey réalisé par le U.S. Census Bureau, puis nous examinons leur performance dans une étude en simulation.

Mots clés : Non-réponse dépendante d'une covariable ; jackknife ; linéarisation ; ajustement par le quotient ; non-réponse uniforme.

1. Introduction

De nombreuses enquêtes-entreprises s'appuient sur un plan d'échantillonnage aléatoire simple stratifié à un degré sans remise. Étant donné l'asymétrie des populations échantillonnées, ces plans comprennent généralement des strates à tirage complet et des strates à tirage partiel. Dans ce genre de plan, les taux d'échantillonnage dans les strates à tirage partiel sont habituellement négligeables (par exemple, moins de 20 % dans toutes les strates). Cependant, si l'unité d'échantillonnage finale est une grande entité commerciale, telle qu'une entreprise, la taille de l'univers est nettement plus petite et, souvent, les fractions d'échantillonnage ne devraient pas être ignorées dans le calcul des estimations de

variance. Des cas de non-réponse se produisent dans la plupart des enquêtes. Nous considérons ici les enquêtes où l'on recourt à l'ajustement de la pondération pour corriger la non-réponse. L'erreur d'échantillonnage n'existant pas dans les strates à tirage complet, les formules de variance classiques ne comprennent aucune composante pour ces strates. En revanche, même dans une strate à tirage complet, la présence d'une non-réponse produit une erreur d'estimation qui est souvent une composante appréciable de l'erreur d'estimation totale.

L'objectif du présent exposé est d'élaborer des méthodes d'estimation de la variance qui tiennent compte de l'ajustement de la pondération pour corriger la non-réponse et de l'existence de strates à tirage complet. À la section 2, nous présentons la notation et les hypothèses, puis nous montrons que les estimateurs de variance obtenus par le jackknife et par linéarisation en ignorant la non-réponse dans les strates à tirage complet, qui sont utilisés à l'heure actuelle dans de

nombreuses enquêtes, sous-estiment la variance réelle de l'estimation repondérée du total de population. En établissant directement une formule de variance approximative, nous obtenons deux estimateurs de variance convergents. Ces estimateurs sont également convergents s'il existe des strates à tirage partiel avec fraction d'échantillonnage importante. Nous élaborons également un estimateur de variance jackknife modifié tenant compte de la variabilité due à la non-réponse dans les strates à tirage complet. À la section 3, nous comparons les estimateurs de variance en utilisant des données couvrant une période de cinq ans provenant de l'Annual Capital Expenditures Survey (ACES) réalisée par le U.S. Census Bureau. À la section 4, nous présentons les résultats de simulations effectuées en utilisant une population créée à partir des données de l'ACES de 2003. Nos résultats de simulation montrent que les estimateurs de variance dans lesquels il n'est pas tenu compte des strates à tirage complet présentent un biais négatif important ; les estimateurs de variance convergents dérivés donnent de bons résultats si les échantillons de strates sont tous de grande taille, mais produisent des résultats non convergents autrement ; et l'estimateur de variance jackknife ne tenant compte d'aucune des fractions d'échantillonnage produit des surestimations. Enfin, à la section 5, nous présentons certaines conclusions.

2. Principaux résultats

Considérons un échantillon stratifié sans remise d'une population finie contenant H strates. Soit n_h et N_h les tailles d'échantillon et de population de la strate h , respectivement, y_{hi} une variable d'intérêt pour laquelle il peut

- Moura, F.A.S., et Migon, H.S. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modelling: An International Journal*, 2, 3, 183-201.
- Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique*, 70, 1, 125-143.
- Pfeffermann, D., Moura, F.A.S. et Silva, P.L.N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 4, 943-959.
- Pfeffermann, D., et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of American Statistical Association*, 102, 1427-1439.
- Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.
- Souza, D.F. (2004). Estimacão de População em Nível Municipal via Modelos Hierárquicos e Espaciais. Unpublished master's dissertation. Universidade Federal do Rio de Janeiro.
- Spiegelhalter, D.J., Thomas, A., Best, N. et Lunn, D. (2004). WinBUGS User Manual Version 1.4. MRC Biostatistics Unit, Cambridge.

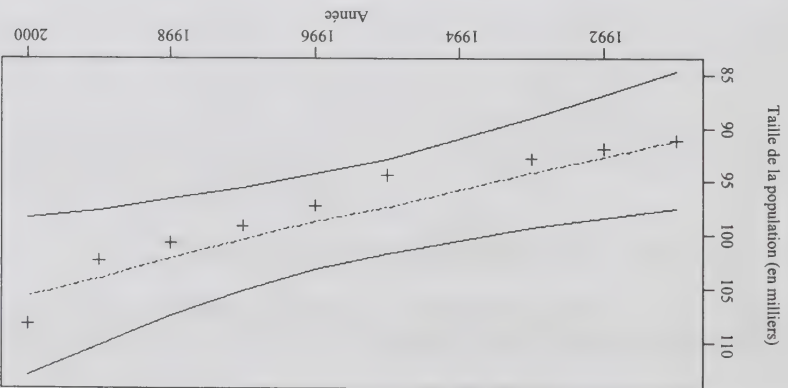


Figure 8 Comparaison entre les tailles de population prédites par le modèle spatial et les statistiques officielles (+) pour une municipalité échantillonnée

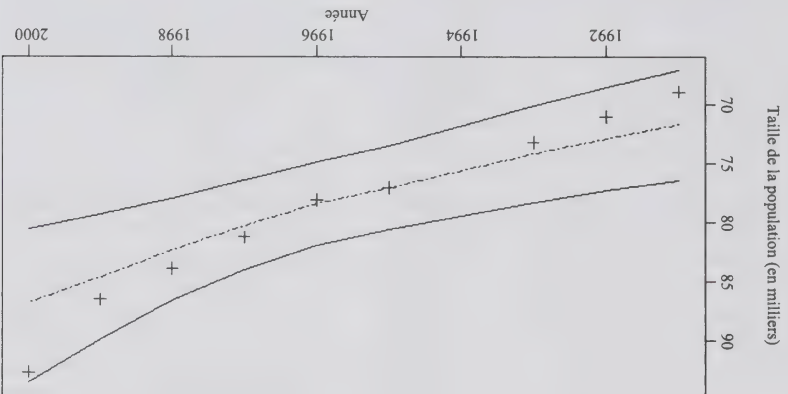


Figure 9 Comparaison entre les tailles de population prédites par le modèle spatial et les statistiques officielles (+) pour une municipalité non échantillonnée

Remerciements

Les auteurs remercient un rédacteur associé et deux examinateurs de leurs commentaires et suggestions constructifs. Les travaux de Fernando Moura et Helio Migon ont été financés en partie par une subvention de recherche du Conseil national brésilien pour le développement de la science et de la technologie (CNPq).

Bibliographie

Besag, J., et Koopman, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.

- Brooks, S.P., et Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 4, 434-455.
- Gelfand, A.E., et Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85, 1, 1-11.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 3, 515-533.
- Migon, H., et Gamerman, D. (1993). Generalized exponential growth models: A bayesian approach. *Journal of Forecasting*, 12, 573-584.
- Mollié, A. (1996). Bayesian mapping of disease. Markov Monte Carlo in Practice. New York : Chapman & Hall, 359-379.
- Richardson, S., Spiegelhalter, D.J., Markov Monte Carlo in Statistics Canada, N° 12-001-X au catalogue

Figure 7 Comparaison entre les tailles de population prédites par le modèle spatial et les statistiques officielles (+) pour la région métropolitaine

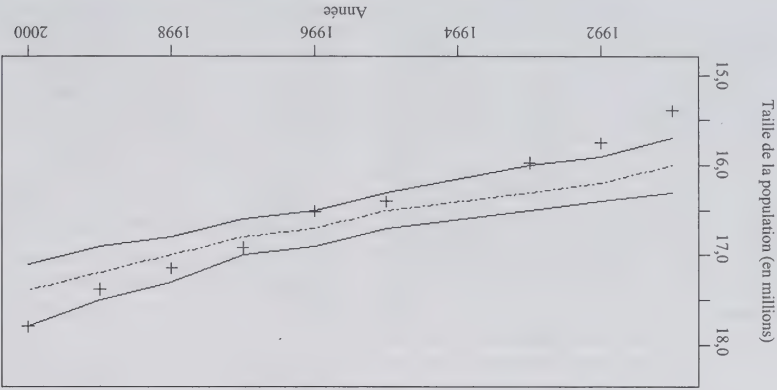
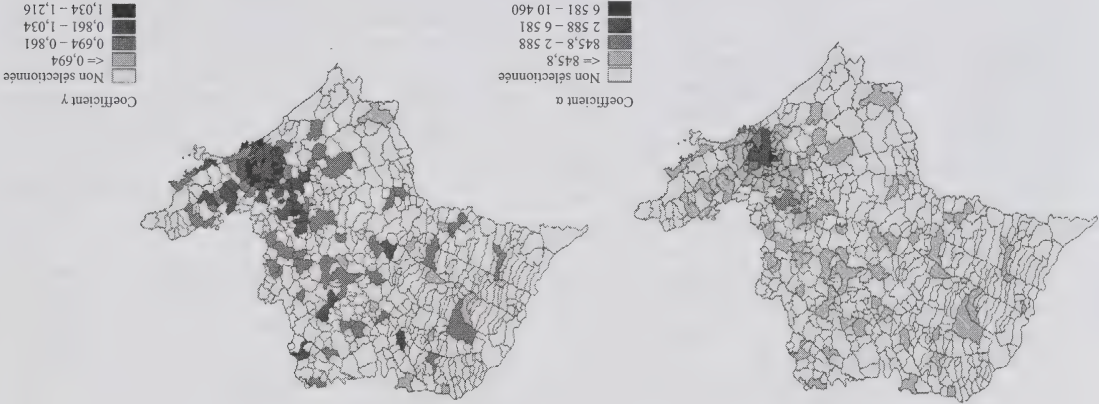


Figure 6 Moyennes a posteriori des paramètres α et γ obtenus au moyen du modèle hiérarchique



Le modèle utilisé dans le présent article permet de dégager la tendance de croissance de la population des municipalités. Des estimations raisonnables des populations municipales sont obtenues pour les années pour lesquelles il existe des données d'enquête, de même que pour celles pour lesquelles on dispose de données de recensement. Les estimations ponctuelles ont une bonne précision et concordent raisonnablement avec les estimations obtenues pour les domaines plus grands en utilisant d'autres techniques. L'information antérieure peut être mise à jour aussitôt que des estimations fondées sur un nouveau

5. Remarques finales

les prédictions obtenues au moyen du modèle spatial concordent raisonnablement avec les chiffres officiels.

D'autres travaux devraient être effectués afin de tenir compte de l'autocorrélation des paramètres d'intérêt au cours du temps. Les renseignements supplémentaires sur les estimations de la variance d'échantillonnage des estimateurs directs pourraient également être considérés comme des données additionnelles. L'hypothèse selon laquelle l'erreur de couverture au recensement est distribuée symétriquement autour de zéro pourrait être relâchée en lui appliquant une distribution non symétrique. Il faut néanmoins pour cela bien connaître la forme de la distribution, ce qui pourrait être difficile en pratique.

recensement ou une nouvelle enquête deviennent disponibles. De surcroît, l'approche proposée fournit la densité de probabilité de la quantité d'intérêt, ce qui facilite le processus de prise de décision.

La figure 6 montre que les moyennes a posteriori des paramètres α et γ qui indiquent le modèle hiérarchique semble être distribuées spatialement. Les paramètres des domaines voisins paraissent plus semblables que ceux des domaines éloignés, ce qui suggère d'appliquer le modèle spatial.

4.5 Choix du modèle

Nous nous sommes servi de la mesure de l'écart par rapport à la prédiction attendue (EPA) (Gelfand et Ghosh 1998) pour choisir le modèle le plus approprié. La mesure de l'EPA est égale à la somme de deux termes. Le premier, désigné par G , peut être interprété comme une mesure de la qualité de l'ajustement et le deuxième, désigné par P , est un terme de pénalité pour les modèles sous-ajustés ainsi que surajustés. Les expressions pour G et P sont données, respectivement, par $G = \sum_{i=1}^m \sum_{j=1}^n (y_{ij}^{hp} - E(y_{ij}^{hp} | M))^2$ et $P = \sum_{i=1}^m \sum_{j=1}^n V(y_{ij}^{hp} | M)$, où les espérances et les variances sont calculées par rapport à la loi prédictive a posteriori associée à une future observation (y_{ij}^{hp}) de y_{ij}^{hp} gérée sous le modèle hypothétique (M). Selon ce critère, le modèle est d'autant meilleur que la valeur est petite. Comme le montre le tableau 6, le critère EPA favorise légèrement le modèle spatial.

4.6 Analyse des résultats

Le niveau le plus agréé pour lequel la PNAD fournit des estimations précises est la région métropolitaine, qui correspond à un ensemble de municipalités contigües. Afin de valider les résultats obtenus au moyen du modèle spatial, nous avons comparé des estimations de population pour la grande région métropolitaine de São Paulo aux projections statistiques officielles. La loi a posteriori de $\mu_i = \sum_{j=1}^n \pi_{ij}^* A_j$ s'obtient facilement en ajoutant $\mu_i^{(t)} = \sum_{j=1}^n \pi_{ij}^{(t)} * A_j$ à l'algorithme MCMC, où μ_i représente la population totale de la région métropolitaine à la période t et r est le nombre de municipalités appartenant à cette région métropolitaine.

Modèle	G	P	EPA
Hierarchique	1,37E+09	6,14E+09	7,51E+09
Spatial	1,05E+09	6,19E+09	7,24E+09

À la figure 7, nous comparons les estimations de population (μ_i) de la région métropolitaine de São Paulo obtenues au moyen du modèle spatial aux statistiques officielles. Les courbes en trait plein représentent les limites des intervalles de crédibilité à 95 % de μ_i , tandis que la courbe en trait interrompu représente les estimations ponctuelles respectives. Le symbole (+) représente les statistiques officielles observées. Soulignons que certaines

projections statistiques officielles se situent en dehors de la limite inférieure de crédibilité (γ compris pour le recensement de 1991). Une enquête plus approfondie devrait donc être faite afin de découvrir les raisons de ces divergences. Cependant, si nous raisonnons la comparaison au niveau des municipalités, la conclusion générale est que les prédictions du modèle et les statistiques officielles concordent raisonnablement. Les intervalles de crédibilité à 95 % contiennent 92,4 % des projections statistiques officielles. La moyenne de l'erreur relative absolue (ERA) entre la densité de population estimée et les projections statistiques officielles est de 3 %. Ces mesures de l'EPA sont, en moyenne, presque les mêmes pour les municipalités sélectionnées et non sélectionnées.

À la figure 8 nous comparons les estimations ponctuelles des tailles de population (μ_i) aux projections statistiques officielles et aux tailles de population officielles selon le recensement pour une municipalité échantillonnée. La méthode de calcul des projections officielles s'appuie sur l'hypothèse qu'un ensemble de petits domaines et un domaine plus grand, qui les contient, ont la même coupe de croissances démographique. La population du domaine plus grand est projetée par la méthode des composantes, puis est répartie proportionnellement entre les petits domaines. La méthode des composantes s'appuie sur des données provenant du recensement le plus récent, ainsi sur les nombres de naissances et de décès et les chiffres de migration nette tirés des dossiers administratifs. La méthode des composantes consiste à projeter la population pour une période t en ajoutant à la population durant une période antérieure le nombre de naissances et le chiffre de migration nette, et en soustrayant le nombre de décès survenus durant le même intervalle de temps.

Les courbes en trait plein représentent les limites des intervalles de crédibilité à 95 % pour μ_i obtenues au moyen du modèle spatial, tandis que la droite en trait interrompu montre les moyennes a posteriori respectives. Le symbole (+) représente la projection de population officielle pour la période intercensitaire et la population observée durant les années de recensement. Il convient de souligner que les estimations ponctuelles sont relativement proches des statistiques projetées officielles et du chiffre de population obtenu l'année du recensement. L'utilisation du modèle proposé semble donc produire des estimations fiables au niveau de la municipalité, avec l'avantage supplémentaire de fournir une mesure de l'erreur respectivement. Nous analysons aussi les estimations obtenues pour certaines municipalités non échantillonnées dans la PNAD. La figure 9 donne les prédictions du modèle, les intervalles de crédibilité à 95 %, les statistiques projetées officielles et les valeurs de population observées aux recensements pour une municipalité non échantillonnée (+). Nous voyons que

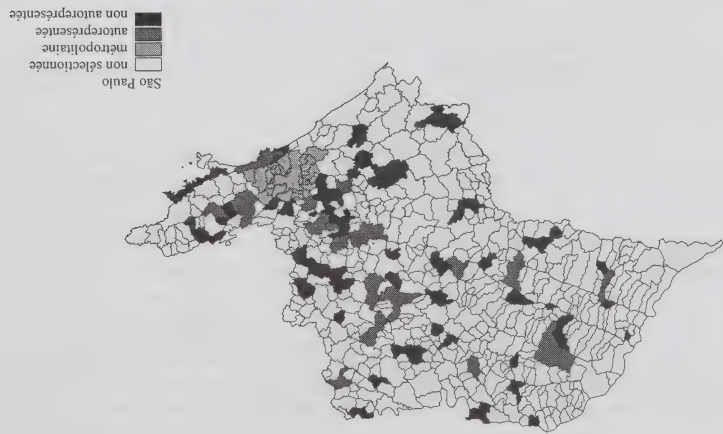


Figure 4 Municipalités de São Paulo échantillonnées pour la PNAD, classifiées selon la définition d'échantillonnage

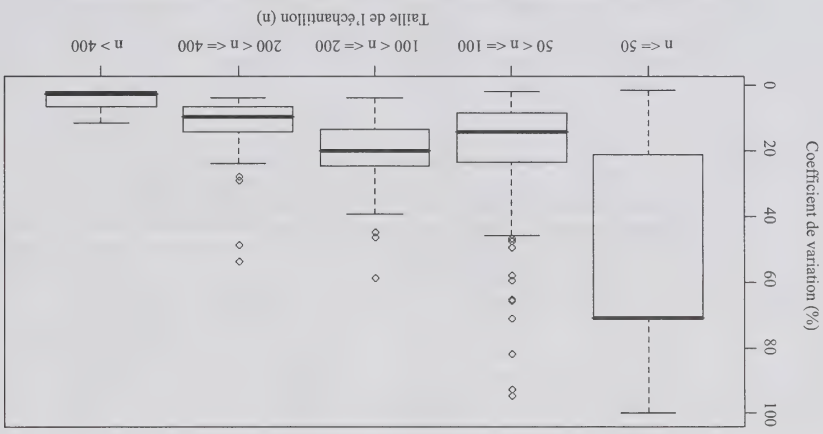


Figure 5 Boîtes à moustaches des coefficients de variation des estimations de population directe

4.4 Certains résultats

Nous avons généré 20 000 échantillons après avoir écarté les 5 000 premiers. Nous ne dégageons aucune preuve de non-convergence des paramètres des modèles hiérarchique et spatial. Une analyse minutieuse des données de sortie MCMC donne à penser que la convergence a été atteinte pour tous les paramètres des modèles. Nous résumons les résultats obtenus par ajustement du modèle hiérarchique (3) aux données fournies par l'enquête PNAD. Nous avons utilisé les moyennes a posteriori des paramètres du modèle comme estimations ponctuelles. Le tableau 5 donne ces estimations ainsi que les racines carrées respectives de la variance a posteriori. L'examen du tableau montre que l'estimation de η_1 est significativement positive, ce qui

concorde avec le résultat attendu selon l'équation 4 : plus la taille d'échantillon est grande, plus σ_n^2 est petite.

Tableau 5
Résumé des lois a posteriori des paramètres du modèle (2)

Paramètre	Moyenne a posteriori	E.-T. a posteriori
α	892,500	202,000
β	105,700	1,278
γ	0,072	0,008
η_0	10,620	0,133
η_1	3,185	0,484
τ_2^2	2,174E-7	2,961E-8
τ_1^2	139,000	19,560

4.3 Spécification des lois a priori

Pour assigner la moyenne des lois a priori normale des paramètres α , β et γ , reliés à l'évolution de la population, nous avons d'abord développé la fonction $\alpha + \beta \exp(\gamma t)$ autour de zéro par un développement en série de Taylor jusqu'au deuxième ordre, puis nous avons égalé l'expression résultante aux valeurs de la densité moyenne aux recensements de 1991 et 2000, ainsi qu'au dénombrement de la population de 1996. En l'absence d'information a priori, nous avons considéré une valeur raisonnablement grande (10^6) pour les variances a priori de α , β et γ .

Donc, nous avons posé $\alpha \sim U(-\infty, +\infty)$ (voir la section 3.3 pour plus de renseignements) pour le modèle spatial et $\alpha \sim N(370, 10^6)$ pour le modèle hiérarchique, ainsi que $\beta \sim N(726, 10^6)$, $\gamma \sim N(0, 04, 10^6)$ pour les deux modèles. Cet ajustement a pour but d'obtenir une valeur raisonnable, mais essentiellement vague, des moyennes a priori. En ce qui concerne les précisions et η_0 , η_1 , nous avons attribué des priors relativement vagues : $\tau_a^2 \sim \text{Ga}(0,001, 0,001)$, $\tau_\gamma^2 \sim \text{Ga}(0,001, 0,001)$, $\eta_0 \sim N(0, 10^6)$ et $\eta_1 \sim N(0, 10^6)$.

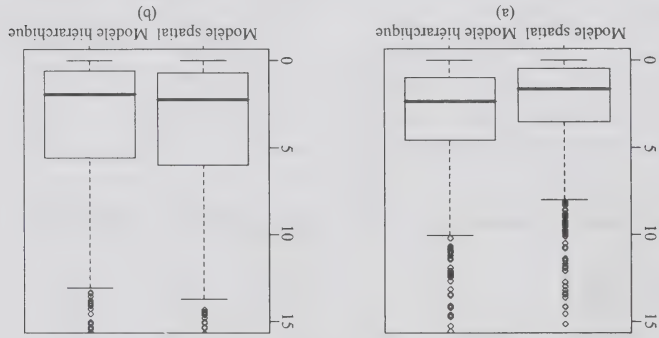


Figure 2 Boîtes à moustaches de l'erreur quadratique moyenne (EQM) pour les cas suivants : a) données générées au moyen du modèle spatial auxquelles sont ajustés respectivement les modèles spatial et hiérarchique et b) données générées au moyen du modèle hiérarchique auxquelles sont ajustés respectivement les modèles spatial et hiérarchique

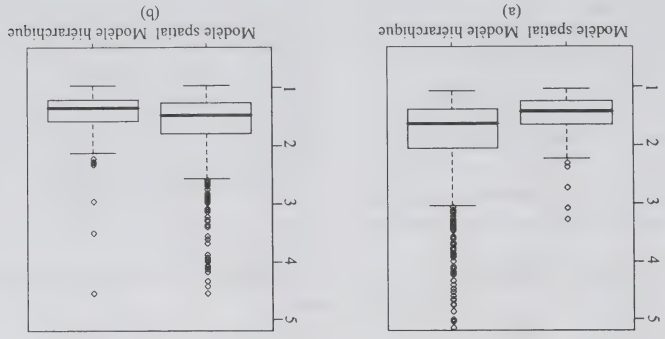


Figure 3 Boîtes à moustaches de l'erreur relative absolue (EKA) pour les cas suivants : a) données générées au moyen du modèle spatial auxquelles sont ajustés respectivement les modèles spatial et hiérarchique et b) données générées au moyen du modèle hiérarchique auxquelles sont ajustés respectivement les modèles spatial et hiérarchique

4. Application

Nous présentons maintenant deux applications de notre approche, la première en nous servant d'un ensemble de données simulées et la deuxième, de l'ensemble de données réelles qui a motivé la présente étude. L'étude par simulation a pour but de vérifier si les paramètres d'intérêt sont estimés correctement et de procéder à une analyse de sensibilité en ce qui concerne la forme des lois a priori utilisées pour ajuster le modèle.

4.1 Application aux données simulées

Nous avons exécuté une petite étude en simulation de l'ajustement des modèles hiérarchique et spatial présentés à la section 3. Nous avons fixé les hyperparamètres réels du modèle reliés à la courbe de croissance à $\alpha = 40$, $\beta = 25$, $\gamma = 0,05$. Donc, nous considérons une situation où la taille de la population double approximativement en 25 ans. Nous avons fixé les paramètres associés au modèle de la variance d'échantillonnage à $\eta_0 = 6,5$, $\eta_1 = 0,5$. Enfin, nous avons fixé les paramètres de précision à $\tau_a^2 = 0,0001$ et $\tau_y^2 = 400$, respectivement. Nous avons fixé les précisions τ_a^2 et τ_y^2 de façon qu'elles concordent avec les échelles des quantités qu'elles mesurent respectivement. La variation relative de l'ordonnée à l'origine entre les domaines est plus importante que celle du paramètre de croissance, ce à quoi nous nous attendons dans les situations pratiques.

Puisqu'il est généralement reconnu que la forme des priors a plus d'incidence sur la composante des paramètres de variance que les paramètres fixes, nous avons ajusté le modèle aux données simulées en utilisant deux priors vagues pour les paramètres liés à la variance : pour l'écart-type, nous avons choisi la loi uniforme, qui est l'un des priors recommandés par Gelman (2006) pour les modèles hiérarchiques linéaires, et pour la précision, la loi gamma, fréquemment utilisée comme loi par défaut dans certains logiciels. Dans le premier cas, nous avons assigné $\sigma_a \sim U(0, 1,000)$ et $\sigma_y \sim U(0, 100)$, où $\sigma_a = 1/\tau_a$ et $\sigma_y = 1/\tau_y$. Dans le deuxième cas, nous avons considéré $\tau_a^2 \sim G(0,001, 0,001)$ et $\tau_y^2 \sim G(0,001, 0,001)$. Pour les autres paramètres, nous avons fixé $\alpha \sim U(-\infty, +\infty)$ pour le modèle spatial (voir la section 3.3 pour plus de précisions) et $\alpha \sim N(0, 10^3)$ pour le modèle hiérarchique. Pour les autres paramètres, nous avons fixé $\beta \sim N(0, 10^3)$, $\gamma \sim N(0, 10^3)$, $\eta_0 \sim N(0, 10^4)$ et $\eta_1 \sim N(0, 10^4)$ pour les deux modèles. Nous avons aussi étudié l'effet du nombre de petits domaines. Nous avons simulé des données distinctes provenant des modèles hiérarchique et spatial avec $m = 60$ et $m = 100$ domaines dans chaque cas. Pour chaque combinaison du nombre de domaines et du modèle employé, nous avons généré 200 ensembles de données. Par conséquent, nous avons simulé en tout 800 ensembles de données artificielles. La distribution des tailles d'échantillon

dans les domaines est la même pour les ensembles de données simulés comptant 60 et 100 domaines. Le tableau 2 donne les fréquences relatives des tailles d'échantillon de petit domaine pour les deux ensembles de données simulés. Ces tailles d'échantillon sont fort semblables à celles provenant des données réelles qui sous-tendent cette étude par simulation. Le nombre de voisins utilisés dans le modèle spatial varie de 1 à 12, et chaque domaine possède en moyenne cinq voisins. Nous avons considéré une période totale de $n = 9$ années.

Tableau 2

Fréquences relatives des tailles d'échantillon de petit domaine pour les deux ensembles de données simulés

Taille d'échantillon	Fréquence relative
2	0,05
3	0,20
4	0,25
5	0,25
6	0,20
7	0,05
8	0,05
9	0,05
10	0,05
11	0,05
12	0,05
13	0,05

Afin d'éliminer la corrélation entre les chaînes, nous avons généré 20 000 échantillons après avoir écarté les 10 000 premiers. Nous ne dégageons aucune preuve de non-convergence des paramètres des modèles hiérarchique et spatial. Une analyse minutieuse de certaines données de sortie provenant des échantillons MCMC pour certains ensembles de simulation donne à penser que la convergence a été atteinte pour tous les paramètres du modèle. Nous avons évalué les propriétés statistiques des estimations de la densité de population (π_n) en examinant l'erreur relative absolue moyenne (ERA) des estimations et l'erreur quadratique moyenne (EQM), respectivement dont données par :

$$ERA_{\pi_n} = \frac{1}{200} \sum_{i=1}^{l=200} \frac{\pi_n^{(i)}}{|\pi_n^{(i)} - \pi_n^{(l)}|}$$
$$EQM_{\pi_n} = \frac{1}{200} \sum_{i=1}^{l=200} (\pi_n^{(i)} - \pi_n^{(l)})^2$$

$l = 1, \dots, m$, $i = 1, \dots, n$. Si l'on s'en tient aux valeurs de l'ERA, la variation est faible. Pour les deux modèles ajustés et les deux tailles d'échantillon de petit domaine testées, les valeurs de l'ERA sont de l'ordre de 1,5 %. Le tableau 3 résume les valeurs de l'EQM obtenues dans les simulations exécutées sous les modèles spatial et hiérarchique avec 60 et 100 domaines et en assignant respectivement un prior gamma et un prior uniforme à la précision et à l'écart-type des paramètres associés à la variance. L'examen du tableau 3 révèle que les EQM ne sont pas affectées par l'utilisation de divers priors vagues. Il convient de souligner que l'accroissement du nombre de domaines, pour passer de 60 à 100, réduit légèrement, de 6 %, la médiane de l'EQM pour le modèle spatial. Par contre, dans le cas du modèle hiérarchique, la diminution est d'environ 13 %.

3.6 Problèmes de calcul

de la fonction d'auto-corrélation permet de déterminer si l'échantillon peut être considéré comme indépendant.

En plus de ces vérifications informelles, nous avons appliqué d'autres critères plus formels. Le critère introduit par Brooks et Gelman (1998) et implémenté dans WinBUGS 1.4 (Spiegelhalter, Thomas, Best and Lunn 2004) permet de diagnostiquer si la dispersion à l'intérieur des chaînes est plus grande que la dispersion entre les chaînes. Considérons I chaînes parallèles et un paramètre d'intérêt λ . Soit λ_i^j la j^{e} valeur de la i^{e} chaîne, pour $i = 1, \dots, K$ et $j = 1, \dots, J$. Les variances entre les chaînes B et à l'intérieur des chaînes W sont alors données par

$$B = J(K-1)^{-1} \sum_{i=1}^K (\bar{\lambda}_i - \bar{\lambda})^2$$

et

$$W = \{K(J-1)^{-1} \sum_{j=1}^J (\lambda_i^j - \bar{\lambda}_i)^2\}$$

où $\bar{\lambda}_i$ et $\bar{\lambda}$ sont, respectivement, la moyenne des observations de la chaîne i , $i = 1, \dots, K$ et la moyenne globale. Sous convergence, ces KJ valeurs sont toutes tirées de la loi a posteriori de λ et la variance de λ peut être estimée de manière convergente par B , W et la moyenne pondérée $\hat{\sigma}_\lambda^2 = (1-1/J)W + (1/J)B$.

Si les chaînes n'ont pas encore convergé, les valeurs initiales influenceront encore l'état stationnaire et $\hat{\sigma}_\lambda^2$ surestimerá σ_λ^2 jusqu'à ce qu'un état stationnaire soit atteint. Par ailleurs, avant la convergence, W aura tendance à sous-estimer σ_λ^2 . En suivant ce raisonnement, Brooks et Gelman (1998) ont proposé une approche graphique itérée qui est implémentée dans WinBUGS 1.4. Elle permet de vérifier si (i) la variance a posteriori pondérée estimée $\hat{\sigma}_\lambda^2$ et la variance à l'intérieur des chaînes W se stabilisent sous forme d'une fonction de J et (ii) le facteur de réduction de la variance, $R = \hat{\sigma}_\lambda^2/W$, s'approche de 1.

Les lois a posteriori des paramètres des modèles proposés ne peuvent pas être exprimées sous une forme explicite. Il est donc nécessaire de recourir à des méthodes d'approximation numérique. Une option, souvent utilisée et facile à appliquer, consiste à produire des échantillons de ces lois en se servant de l'algorithme Monte Carlo par chaînes de Markov (MCMC). Puisque les lois conditionnelles complètes de tous les paramètres des modèles possèdent une forme explicite, sauf pour le vecteur $\gamma = (\gamma_1, \dots, \gamma_K)$, nous nous sommes servi de l'algorithme de l'échantillonneur de Gibbs avec une étape de l'algorithme d'acceptation-rejet pour l'échantillonnage à partir du vecteur γ . Soit π_n la densité de population dans le t^{e} domaine à la période t . Les étapes qui suivent résument comment tirer des échantillons de la loi a posteriori de π_n :

1. Produire $\alpha_{(t)}^{(i)}, \beta_{(t)}^{(i)}, \gamma_{(t)}^{(i)}, \alpha_{(t)}^{(o)}, \gamma_{(t)}^{(o)}, r_{2(o)}^{(i)}, r_{2(o)}^{(o)}, r_{1(o)}^{(i)}, r_{1(o)}^{(o)}$ et $\eta_{(t)}^{(i)}$ pour $i = 1, \dots, M$, où M est le nombre d'échantillons générés à partir des lois conditionnelles complètes de tous les paramètres du modèle, y compris les effets aléatoires;
2. calculer $\pi_n^{(i)} = \alpha_{(t)}^{(i)} + \beta_{(t)}^{(i)} \exp(\gamma_{(t)}^{(i)})$;

Durant l'ajustement des modèles que nous proposons, nous avons appliqué trois vérifications informelles, basées sur des techniques graphiques, pour évaluer la convergence. Elles consistent à observer l'histogramme, la trace et la fonction d'auto-corrélation pour chacune des valeurs échantillonnées calculées. L'analyse de l'histogramme nous permet de repérer les écarts éventuels par rapport à la convergence, tel que la présence de modes multiples. La trace des chaînes simulées en parallèle, chacune avec un point de départ différent et surdispersées par rapport à la loi cible, donne une idée grossière du comportement stationnaire quand les suites de valeurs ont tendance à osciller dans la même région. La représentation graphique

Tableau 1

Sommaire des modèles utilisés

Modèle	Paramètres	Variances	Loi a priori
Hérarchique	$\alpha_i = \alpha + \varepsilon_{\alpha_i}$ β $\gamma_i = \gamma + \varepsilon_{\gamma_i}$	$\log(\sigma_{\eta_i}^2) = \eta_0 + \eta_1(1/\eta_i)$ pour les données d'enquête $\sigma_{\eta_i}^2$ est supposée connue	$\eta_0 \sim N(\eta_0, \phi_{\eta_0})$ $\eta_1 \sim N(\eta_1, \phi_{\eta_1})$
Spatial	$\alpha_i = \alpha + \delta_{\alpha_i}$ β $\gamma_i = \gamma + \varepsilon_{\gamma_i}$	$\log(\sigma_{\eta_i}^2) = \eta_0 + \eta_1(1/\eta_i)$ pour les données de recensement $\sigma_{\eta_i}^2$ est supposée connue	$\sum_{i=1}^m \delta_{\alpha_i} = 0$ $\delta_{\alpha_i} \delta_{\alpha_{i-1}}, \tau_{\alpha}^2 \sim N(\delta_{\alpha_i}, \tau_{\alpha}^2/w_{i+})$

La figure 1 montre les densités démographiques des municipalités de São Paulo en 1991. Ces municipalités ont tendance à être concentrées géographiquement en fonction de classes de densité, ce qui donne à penser que l'application du modèle spatial peut être fructueuse.

3.4 Modélisation des variances d'échantillonnage

Comme nous utilisons des données provenant de deux sources différentes, il est logique de supposer que les variances d'échantillonnage varient au cours du temps. En outre, nous pouvons également considérer que les variances varient selon le domaine.

Dans le cas des années pour lesquelles des données sont fournies par la PNAD, nous supposons que les variances d'échantillonnage sont données par le modèle suivant :

$$\log(\sigma_n^2) = \eta_0 + \eta_1 \cdot (1/n_i) \quad (4)$$

avec n_i représentant le nombre de secteurs de dénombrement échantillonnés dans le i^{e} domaine. Ce modèle traduit l'espérance que la variance diminue à mesure que la taille de l'échantillon augmente.

Pour les années durant lesquelles les recensements ont été exécutés, nous supposons que σ_n^2 est connue et que $\log(\sigma_n^2) = \log(v_n)$ où v_n est calculée de telle manière que l'erreur de couverture au recensement soit égale à 5 % pour tous les domaines. Cette hypothèse implique que, dans



Figure 1 Densité de population des municipalités de São Paulo en 1991

3.5 Sommaire des modèles

Les lois a priori des paramètres communs des modèles spatial et hiérarchique sont les mêmes que celles déjà décrites pour le premier. Les lois suivies par les effets spatiaux aléatoires sont spécifiées à la section 3.3. La variance σ_n^2 a été énoncée de la même façon dans le modèle spatial que dans le modèle hiérarchique. Le tableau 1 résume les modèles utilisés à la section 4. Pour simplifier, nous avons exécuté l'application en fixant $\phi = 1$ dans les deux modèles.

chaque domaine pour les années de recensement, la population réelle est comprise dans l'intervalle donné par la population observée au recensement plus ou moins 5 % de cette valeur. Par conséquent, pour les années de recensement, nous fixons l'écart-type à $\sigma_n^2 = 0,05 \cdot (y_n^2/2)$. Supposer que la variance est connue pour les années de recensement est un moyen d'accorder plus de poids aux données de recensement complet fournissant des renseignements plus fiables que des données d'enquête. Nous supposons que les paramètres η_0 et η_1 suivent des lois normales indépendantes : $\eta_k \sim N(\mu_{\eta_k}, \phi_{\eta_k})$; $k = 0, 1$. Afin d'attribuer des priors vagues aux η_i , nous donnons, pour chaque prior, une valeur nulle à la moyenne et de grandes valeurs aux ϕ_{η_i} . Voir la section 4.2 pour plus de renseignements.

recensement (voir la section 3.4 ainsi que les remarques

finales à la section 5) :

$$(2) \quad \begin{aligned} \gamma'' &= \pi'' + \varepsilon_{\gamma''}, \varepsilon_{\gamma''} \sim N(0, \sigma_{\gamma''}^2) \\ \pi'' &= \{\alpha_j + \beta \exp(\gamma_j t')\}_{1/\phi} \\ \alpha_j &= \alpha + \varepsilon_{\alpha_j}, \varepsilon_{\alpha_j} \sim N(0, \sigma_{\alpha_j}^2) \\ \gamma_j &= \gamma + \varepsilon_{\gamma_j}, \varepsilon_{\gamma_j} \sim N(0, \sigma_{\gamma_j}^2) \end{aligned}$$

où les lois a priori de α , β et γ sont données par :

$\alpha \sim N(\mu_{\alpha}, \sigma_{\alpha}^2)$, $\beta \sim N(\mu_{\beta}, \sigma_{\beta}^2)$, $\gamma \sim N(\mu_{\gamma}, \sigma_{\gamma}^2)$. Il convient

de souligner que l'information provenant de tous les

domaines est obtenue au moyen de la structure hiérarchique

des paramètres α_j et γ_j . Une autre façon de permettre

l'empunt d'information entre les municipalités consiste à

supposer que les α_j sont spatialement structurés (voir la

section 3.3). Si nous supposons que la moyenne π'' est non

explosive, nous pouvons considérer le paramètre $\alpha^{1/\phi}$

comme étant la valeur à laquelle la population municipale

l'évolution de la densité de population au cours du temps.

Les lois a priori de α , β et γ peuvent être choisies en

profitant d'une certaine connaissance démographique

a priori de l'évolution prévue de la population. Dans notre

application, nous fixons $\phi = 1$, ce qui implique que, pour

$t = 0$, la valeur réelle de la densité de population dans

chaque municipalité est donnée par $\alpha_j + \beta$. La structure

hiérarchique impose aux paramètres α_j l'implication que la

valeur espérée de la densité réelle pour toute municipalité à

la période $t = 0$ est $\alpha + \beta$. Supposer que les paramètres de

croissance, γ_j , possèdent une structure hiérarchique signifie

que les densités ont des taux de croissance différents, mais

ont en commun la même moyenne. Une petite étude en

simulation (voir la section 4.1) nous dicte de maintenir le

paramètre β fixe pour tous les domaines, sans aucune perte

de généralité, puisque les niveaux diffèrent encore pour

différentes municipalités. Dans tous les modèles considérés

dans notre application, nous supposons que $\tau_2^2 = \sigma_{\tau_2}^2 \sim$

$G(a_{\tau_2}, b_{\tau_2})$, $\tau_1^2 = \sigma_{\tau_1}^2 \sim G(a_{\tau_1}, b_{\tau_1})$. Afin d'attribuer des priors

vagues, à la section 4.2, nous donnons des valeurs faibles

aux paramètres reliés à ces lois a priori de la précision.

L'hypothèse selon laquelle la fonction moyenne π'' est

donnée par une courbe de croissance exponentielle permet

d'effectuer une correction en cas d'accroissement ou de

diminution de la densité de la population. Les sources de

données utilisées ont des périodes de référence différentes et

les données ne sont pas réparties de la même façon dans le

temps. Dans chaque cas, l'utilisation d'une courbe de

croissance exponentielle offre un avantage supplémentaire,

puisque nous pouvons simplement produire une échelle de

temps afin de nous conformer aux différentes sources de

données, comme nous l'expliquons dans la description de

l'application à la section 4.

Dans le modèle hiérarchique présenté à la section

précédente, l'information provenant de tous les domaines

est combinée afin de prédire la population d'un domaine

particulier. Cependant, il est raisonnable de supposer que les

densités démographiques de deux municipalités voisines

sont plus semblables que celles de deux autres choisies

arbitrairement. La structure régionale est représentée par la

loi a priori conjointe des effets spatiaux aléatoires. Nous

considérons que deux domaines sont voisins s'ils ont une

limite commune.

Dans le modèle que nous proposons, la densité

démographique dans un domaine i à la période t , π''_i , est

afféctée par les domaines voisins par ajout d'effets spatiaux

aléatoires δ_{α_j} aux paramètres α_j , c'est-à-dire $\alpha_j =$

$\alpha + \delta_{\alpha_j}$, où α est un terme représentant l'ordonnée à

l'origine. Par conséquent, α_j varie uniquement avec l'effet

spatial, ce qui représente un effet local, tandis que les

paramètres de croissance γ_j sont considérés comme

semblables dans tous les secteurs (effet global).

La relation entre les secteurs (effet global) est

lois a priori des δ_{α_j} . La loi conjointe a priori de $\delta_{\alpha_j} =$

$(\delta_{\alpha_1}, \dots, \delta_{\alpha_m})'$ sachant l'hyperparamètre $\sigma_{\alpha_j}^2$ est définie

comme dans Mollie (1996) :

$$(3) \quad p(\delta_{\alpha} | \sigma_{\alpha}^2) \propto \frac{\sigma_{m/2}^{\alpha}}{1} \exp \left\{ -\frac{1}{2\sigma_{m/2}^2} \sum_{m=1}^m \sum_{k=1}^{k-1} w_{ik}^{\alpha} (\delta_{\alpha_i} - \delta_{\alpha_k})^2 \right\}$$

où w_{ik}^{α} représente les poids associés à la structure régionale.

Les poids sont choisis de manière que $w_{ik}^{\alpha} = 1$, si i et k

sont contigus, et $w_{ik}^{\alpha} = 0$, autrement. La loi de $\delta_{\alpha} | \sigma_{\alpha}^2$ est

évidemment incorrecte, puisque nous pouvons ajouter

n'importe quelle constante à tous les δ_{α_j} , sans que

contrainte pour nous assurer que le modèle est identifiable.

Nous posons que $\sum_{m=1}^m \delta_{\alpha_j} = 0$ et attribuons à l'ordonnée à

l'origine α une loi a priori uniforme sur l'ensemble de la

droite réelle. Il n'est pas difficile de voir que cette procédure

donne lieu à une densité de vraisemblance $(m-1)-$

dimensionnelle appropriée. Voir Besag et Kooperang (1995)

pour plus de renseignements.

La loi a priori conditionnelle de δ_{α_j} , sachant les effets

δ_{α_i} , des secteurs restants et l'hyperparamètre $\sigma_{\alpha_j}^2$, est

normale de moyenne et de variance données par :

$$E[\delta_{\alpha_i} | \delta_{\alpha_j}, k \in \partial i, \sigma_{\alpha_j}^2] = \delta_{\alpha_i}$$

$$\text{Var}[\delta_{\alpha_i} | \delta_{\alpha_j}, k \in \partial i, \sigma_{\alpha_j}^2] = \frac{\sigma_{\alpha_j}^2}{w_{i+}}$$

où δ_{α_i} désigne la moyenne arithmétique des δ_{α_j} , pour

$k \in \partial i$ (les domaines contigus de i), et $w_{i+} = \sum_{m=1}^m w_{ik}^{\alpha}$ est

le nombre de municipalités voisines de i .

l'étude. En ce qui concerne l'inférence pour petits domaines sous échantillonnage informatif, Pfeiffermann et Sverchkov (2007) est une bonne référence. Nous recommandons également de consulter Pfeiffermann, Mouta et Silva (2006) aux lecteurs qui aimeraient savoir comment suivre une approche de modélisation hiérarchique bayésienne sous échantillonnage informatif.

3. Spécification du modèle

3.1 Modèle de croissance exponentielle

Soit y_t les valeurs d'échantillon d'une loi appartenant à une famille exponentielle dont la valeur prévue est donnée par $\pi_t' = E(y_t' | \theta_t')$ où θ_t' est un vecteur de paramètres inconnus.

Une classe importante et vaste de modèles de croissance exponentielle paramétrisés par $(\alpha, \beta, \gamma, \phi)$ est définie par :

$$(1) \quad \pi_t' = [\alpha + \beta \exp(\gamma t)]^{\phi}.$$

Certains cas particuliers bien décrits dans la littérature sont les distributions :

- (1) Logistique : avec $\phi = -1$, $\pi_t' = \alpha + \beta \exp(\gamma t)$;
- (2) Gompertz : avec $\phi = 0$, en définissant (1) comme $\log(\pi_t) = \alpha + \beta \exp(\gamma t)$;
- (3) exponentielle modifiée : avec $\phi = 1$, $\pi_t' = \alpha + \beta \exp(\gamma t)$.

Le principal avantage de l'utilisation du modèle (1) est qu'il est possible de garder l'échelle originale des observations y_t et de ne changer que la trajectoire de π_t , ce qui facilite l'interprétation. De surcroît, les intervalles de temps ne doivent pas être tous de la même longueur, si bien que les données peuvent provenir de diverses sources de référence (voir la section 4 pour plus de renseignements).

Quand $\psi = \exp(\gamma) > 1$, le processus est non explosif, ce qui implique que π_t converge vers $\alpha^{1/\phi}$ quand $t \rightarrow \infty$, sous la condition que, pour $\phi = 0$, cette quantité est égale à $\log(\alpha)$. Quand $\psi > 1$, les courbes sont concaves pour $\phi \geq 0$ et $\beta > 0$, donnant lieu à un processus explosif. Cette classe de modèles est celle des modèles généralisés de croissance exponentielle. Milgrom et Garman (1993) montrent comment le modèle de croissance exponentielle peut être vu comme un cas particulier d'un modèle dynamique général.

3.2 Modèles de croissance hiérarchiques

Dans le présent article, les principaux paramètres d'intérêt π_t'' sont les fonctions de croissance exponentielle non linéaire

raisons d'utiliser les densités.

Pour chaque période, les estimations de ces quantités ne sont disponibles que pour $k < m$ municipalités correspondantes aux unités de premier degré de l'échantillon de la PNAD. Afin d'estimer la densité de population municipale, nous divisons simplement l'estimation de la population totale par la superficie de la municipalité.

Soit y_t'' la densité de population obtenue d'après les données de recensement ou estimées d'après la PNAD à la période t , $t = 1, \dots, n$ pour la i^{e} municipalité, $i = 1, \dots, m$. Notre but est de faire des inférences au sujet de la densité de population réelle π_t'' pour la population de toutes les municipalités, y compris celles qui ne sont pas échantillonnées. À la section suivante, nous modélisons les densités réelles de population municipale π_t'' au moyen d'une fonction de croissance hiérarchique non linéaire stochastique. Nous supposons que les quantités aléatoires y_t'' suivent une loi normale de moyenne π_t'' et de variance σ_t'' .

Nous adoptons une approche bayésienne pour cette étude. Par conséquent, les prédictions sont décrites par des densités de probabilité, ce qui donne aux utilisateurs l'occasion d'analyser les incertitudes que comporte le processus de décision. Ce fait est l'un des avantages, parmi de nombreux autres, de l'utilisation de ce genre d'approche.

Les valeurs de y_t'' ne sont obtenues pour toutes les municipalités de l'État de São Paulo que pour les années de recensement. Bien que l'on s'efforce, durant le recensement, d'obtenir le dénombrement complet de toute la population, des erreurs de couverture peuvent avoir lieu. Par conséquent, nous émettons l'hypothèse du modèle qui suit pour les données de recensement ainsi que pour celles provenant de la PNAD, à l'exception des variances σ_t'' , qui sont fixées à une valeur plus faible pour les données de

consiste à renforcer l'estimation en empruntant de l'information à tous les domaines et à d'autres sources de données connexes. Comme l'a énoncé Pfeffermann (2002), les sources de données appropriées pour cette tâche peuvent être classées en deux catégories, à savoir les données provenant d'autres domaines semblables en ce qui concerne les caractéristiques d'intérêt, les données antérieures obtenues pour la caractéristique d'intérêt et l'information auxiliaire. Dans notre contexte démographique, la principale source de données connexes comprend les recensements de 1991 et de 2000, ainsi qu'un dénombrement complet de la population exécuté en 1996.

Le but de la présente étude est d'obtenir des estimations des populations municipales fondées sur des données d'enquête fournies par la PNAD et sur des données de recensement. Nous proposons un modèle hiérarchiquement non structuré et nous évaluons la qualité de son ajustement et son pouvoir prédictif. Nous envisageons également un modèle hiérarchique spatialement structuré dans l'esprit de Moura et Migon (2002), puisque le chiffre de population par domaine et son profil de croissance pourraient être reliés au développement des domaines voisins. Par souci de simplicité, dans la suite de l'exposé, nous donnons respectivement aux modèles hiérarchique non structuré et hiérarchique structuré spatialement les noms de modèle hiérarchique et modèle spatial.

À la section 2, nous décrivons les principales sources de données utilisées dans nos travaux. À la section 3, nous présentons les modèles proposés, ainsi qu'un critère de sélection de modèle. À la section 4, nous présentons des applications à des données réelles, ainsi qu'à des données simulées. Enfin, à la section 5, nous résumons brièvement l'étude et décrivons dans les grandes lignes nos futurs travaux de recherche.

2. Ensemble de données

Les données d'entrées pour les modèles présentées à la section 3 proviennent des cycles de 1992 à 1999 de l'enquête annuelle sur les ménages (PNAD), des recensements de 1991 et de 2000, et d'un dénombrement complet de la population effectué en 1996. Afin d'évaluer l'approche proposée, nous considérons les municipalités de l'État de São Paulo comme étant les domaines d'intérêt.

À la présente section, nous décrivons brièvement les sources des données, en mentionnant leurs principaux avantages et limites. Nous avons tiré de la PNAD les estimations démographiques directes pour les municipalités échantillonnées. Comme nous l'expliquons à la section 3, ces estimations servent de données d'entrée pour l'inférence au sujet de nos paramètres cibles. Nous utilisons aussi dans

notre application les deux recensements et le dénombrement de population de 1996.

Le recensement démographique du Brésil est la source principale d'information au sujet de la population. Il est effectué tous les dix ans, habituellement au début de la décennie. Bien que l'objectif consiste à compter tous les membres de la population, certains erreurs de recensement sont découvertes. L'importance des erreurs est évaluée au moyen d'une enquête postcensitaire exécutée peu après l'achèvement du recensement.

L'enquête annuelle sur les ménages (PNAD) est conçue pour produire des renseignements de base sur la situation socioéconomique du pays. L'unité étudiée est le ménage, pour lequel sont recueillis des renseignements annuels sur le nombre de membres, leur sexe, leur niveau d'études, leur situation d'emploi, etc. L'enquête n'est pas exécutée durant les années de recensement et n'a pas été réalisée en 1994 pour des raisons administratives. L'échantillon est sélectionné selon un plan d'échantillonnage en grappes à trois degrés. Les unités primaires et secondaires d'échantillonnage sont respectivement la municipalité et le secteur de recensement (qui compte, en moyenne, 250 ménages). Les municipalités sont stratifiées en fonction de la taille de leur population déterminée d'après le dernier recensement. Au premier degré, toutes les municipalités appartenant aux régions métropolitaines et aux capitales des États (lesquelles, au Brésil, sont normalement les plus grandes villes dans les États respectifs) sont échantillonnées. Les municipalités dont la population est supérieure à une certaine valeur seuil sont également incluses dans l'échantillon avec une probabilité de un. Celles qui restent sont stratifiées et deux d'entre elles sont échantillonnées dans chaque strate avec une probabilité proportionnelle à la taille de leur population.

Les secteurs de dénombrement sont échantillonnés avec une probabilité proportionnelle au nombre de ménages résidant dans le secteur au moment du dernier recensement. Enfin, au dernier degré, les ménages sont échantillonnés systématiquement avec une probabilité égale au moyen d'une liste qui est mise à jour au début de l'enquête. Les mêmes municipalités et secteurs de dénombrement sont gardés pour toutes les enquêtes exécutées durant une décennie particulière, tandis que les ménages sont échantillonnés chaque année.

Puisque chaque secteur est échantillonné avec une probabilité proportionnelle à son nombre respectif de ménages, on pourrait soutenir que le mécanisme d'échantillonnage est informatif en ce qui concerne la population du secteur. Toutefois, puisque la variable réponse effective-ment utilisée dans la présente étude est la densité de population par domaine, il est raisonnable de supposer que le mécanisme de sélection de l'échantillon n'est pas pertinent. Donc, nous n'abordons pas cette question dans

Prediction de la population de petits domaines au moyen de modes hierarchiques

Debora F. Souza, Fernando A.S. Moura et Helio S. Migon¹

Resumé

Le présent article décrit une méthode de prédiction pour petits domaines fondée sur des données tirées d'enquêtes périodiques et de recensements. Nous appliquons cette méthode pour obtenir des prédictions démographiques pour les municipalités non échantillonnées dans l'enquête annuelle sur les ménages du Brésil (PNAD), ainsi que pour accroître la précision des estimations fondées sur le plan de sondage obtenues pour les municipalités échantillonnées. En plus des données fournies par la PNAD, nous utilisons des données démographiques provenant des recensements de 1991 et de 2000, ainsi que d'un dénombrement complet de la population effectué en 1996. Nous proposons et comparons des modèles de croissance hiérarchiquement non structurés et spatialement structurés qui gagnent en puissance en s'appuyant sur toutes les municipalités échantillonnées.

Mots clés : Méthode de Monte Carlo par chaînes de Markov (MCMC) ; projection démographique ; modèles spatiaux.

1. Introduction

Comme dans de nombreux autres pays, la demande de statistiques détaillées et à jour sur des petits domaines a augmenté régulièrement au Brésil. La nécessité de brosser un tableau plus précis des sous-régions, en vue de résoudre des problèmes de distribution, d'équité et de disparité, est à l'origine de cet accroissement de la demande. Par exemple, certains sous-régions ou certains sous-groupes pourraient, à certains égards, être à la traîne par rapport à la moyenne globale. Par conséquent, il est nécessaire de repérer ces régions et d'obtenir des renseignements statistiques à ce niveau géographique avant de pouvoir prendre toute mesure éventuelle de correction. Outre ces exigences nationales, les autorités locales doivent disposer d'estimations fiables, comme les censeurs démographiques, à des fins d'analyse, de planification et d'administration.

Au Brésil, un exemple important de demande d'estimations fiables a trait à la façon dont le revenu fédéral, qui doit être partagé en vertu de la constitution, est réparti entre les diverses municipalités (le Brésil est une république fédérée constituée d'États et du District fédéral. Les États sont subdivisés en municipalités, qui partagent des caractéristiques de villes et des comités – elles peuvent contenir plus d'une région urbaine, mais elles ne sont dotées que d'un seul maître et d'un seul conseil municipal). Le nombre prédit d'habitants de la municipalité est utilisé par le gouvernement fédéral comme critère pour affecter les fonds. D'où la nécessité d'obtenir des prévisions fiables de la population municipale afin d'appliquer équitablement ce critère, réglementé par la loi fédérale.

Le problème de l'estimation sur petits domaines a suscité de l'intérêt dans la littérature statistique à cause de la demande croissante d'information statistique détaillée des secteurs public et privé. Un excellent exposé à jour sur les méthodes d'estimation sur petits domaines et leurs applications peut être consulté dans Rao (2003). Les données sur les petits domaines proviennent principalement d'enquêtes périodiques dont les tailles d'échantillon ne sont pas suffisamment grandes pour pouvoir produire des estimations fiables pour les domaines. Un moyen d'aborder le problème

est, en général, ne fournir pas de mesures de l'erreur des estimations. Elle ne tient pas compte de toutes les incertitudes du modèle. Elle ne tient pas compte de l'évolution hypothétique de toutes les municipalités. Le principal inconvénient de cette méthode est qu'elle dépend de l'évolution hypothétique de la mortalité, de migration et de migration pour les mêmes pour les municipalités. À son tour, la prédiction pour une donnée auxiliaire pour répartir la population totale prédite par une plus grande région, puis en utilisant des données démographiques municipales en se basant d'abord sur une méthode courante consiste à obtenir des estimations

ne sont pas échantillonnées du tout. L'approche courante consiste à obtenir des estimations directes. De surcroît, un nombre important de municipalités acceptables quand sont utilisées des estimations par sondage suffisamment grandes pour produire des erreurs-types acceptables, les tailles des échantillons municipaux ne sont pas au niveau municipal. Autrement dit, à part quelques municipalités importantes de données démographiques. Toutefois, l'enquête annuelle sur les ménages (PNAD) est une

1. Debora F. Souza, Département des méthodes et de la qualité, IBGE, Rio de Janeiro, Brésil. Courriel : fmoura@im.ufrj.br; Helio S. Migon, Universidade do Brasil-UFRJ, Rio de Janeiro, Brésil. Courriel : migon@im.ufrj.br; Fernando A.S. Moura, Universidade do Brasil-UFRJ, Rio de Janeiro, Brésil. Courriel : debora.souza@ibge.gov.br

plus faible dans une situation d'enquête par sondage, car l'incertitude de l'estimation résumée par les termes h_{2a} et h_{3a} augmente.

6. Résumé

Nous avons décrit une approche de modélisation mixte double qui étend la méthode GSPREE à l'estimation de la

composition sur petits domaines en présence de données manquantes différentes. Nous avons calculé une EQMCP approximative qui contient une décomposition en trois parties, correspondant à la variance de prédiction de l'effet aléatoire inconnu, la variance d'échantillonnage en l'absence de données manquantes et la variance supplémentaire due aux données manquantes, respectivement. L'approche a été appliquée aux données de registre sur les ménages de la Norvège et a donné des corrections utiles pour tenir compte des numéros d'identification du logement manquants informatifs.

Remerciements

L'auteur remercie le rédacteur associé et les examinateurs de leurs commentaires et suggestions qui lui ont permis d'améliorer cet article.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis*. New York : John Wiley & Sons, Inc.

Booth, J.G., et Hoberg, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.

Breslow, N.E., et Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (avec discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 93-273-282.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlini, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Longford, N. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society, Series A*, 162, 227-245.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*. Londres : Chapman and Hall.

Prasad, N.G., et Rao, J.N.K. (1990). The estimation of mean square errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Purcell, N.J., et Kish, L. (1980). Postcensal estimates for local areas (or domains). *Revue Internationale de Statistique*, 48, 3-18.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.

Zhang, L.-C., et Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 479-496.

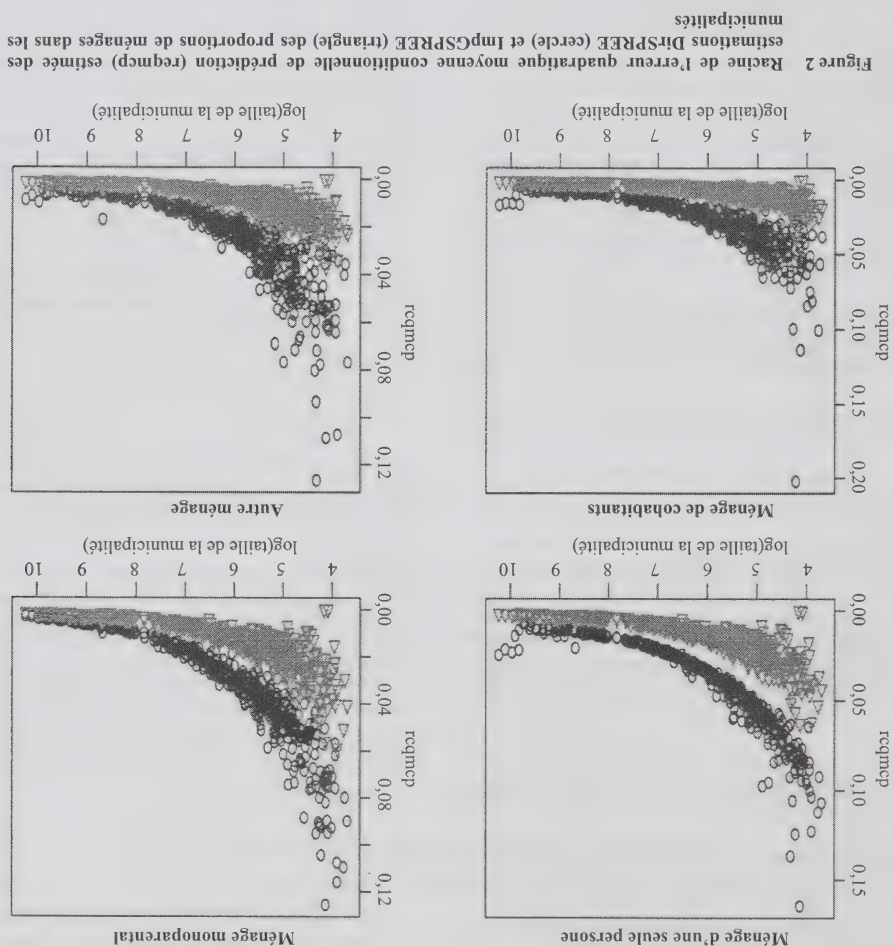


Figure 2 Racine de l'erreur quadratique moyenne conditionnelle de prédiction (reqmcp) estimée des estimations Disproportion (cercle) et ImpGSPREE (triangle) des proportions de ménages dans les municipalités

Ensemble, les expressions mènent à une décomposition en trois parties de l'EQMCP semblable aux expressions (12) à (14). Dans l'estimation de l'EQMCP qui suit, nous ignorons l'effet de l'API. Cela se justifie dans notre cas, parce que l'API équivaut essentiellement à un ajustement multiplicatif constant très proche de l'unité, comme le montre le graphique du milieu à gauche de la figure 1.

L'EQMCP d'un dénombrement Disproportion est calculée comme étant la variance d'«échantillonnage» qui est induite par des données manquant au hasard dans chacune des cellules du tableau à double entrée, plus un terme de biais quadratique qui est estimé par le carré de la différence entre le dénombrement ImpGSPREE et le dénombrement Disproportion correspondant, à condition que l'hypothèse (9) soit un modèle plus approprié pour les données manquantes que l'hypothèse (10).

Les racines estimées de l'EQMCP (reqmcp) sont données à la figure 2. En moyenne, les deux valeurs diminuent à mesure que la taille de la municipalité augmente. Cependant, pour certaines des municipalités les plus grandes, l'EQMCP de la proportion Disproportion est anormalement grande pour les ménages d'une seule personne et les ménages de cohabitants à cause du terme de biais. Dans l'ensemble, l'EQMCP de la composition ImpGSPREE est clairement plus petite que celle de la composition Disproportion. Le terme h_{α} , qui correspond à la variance de prédiction de X_{α} , est de loin la contribution dominante à l'EQMCP (plus de 99 % dans de nombreux domaines). Cela est compréhensible, puisque l'échantillon « observé » contient plus de 550 000 personnes, de sorte que l'incertitude de l'estimation des paramètres est comparativement négligeable. Mais le pourcentage mentionné sera

Pour la quasi-vraisemblance (5), nous supposons que $v_1 = 1$. Soit $t_{ak} = X_{ak}/X_a$. Nous avons

$$V(t_{ak}) = N_a^{-1} \theta_{ak} (1 - \theta_{ak}) X_a^{(2)} / X_a^2$$

et

$$\text{Cov}(t_{ak}, t_{aj}) = -N_a^{-1} \theta_{aj} \theta_{ak} X_a^{(2)} / X_a^2.$$

grands.

Dans les graphiques de la ligne inférieure de la figure 1,

les estimations sont obtenues en utilisant l'approche de modélisation mixte double. Dans le graphique situé en bas à gauche, les estimations sont obtenues par ajustement proportionnel itératif en partant des compositions en super-population estimées $\{\theta_{ak}\}$, désignées par SupGSPREE. L'accroissement postcensitaire extrême dans les municipalités les plus grandes est réduit. Mais les variations par rapport aux proportions fondées sur le recensement sont clairement réduites de manière excessive vers la moyenne de population pour les domaines plus petits. Par exemple, la variation est nettement moindre que celle de $\theta_{ak}^* - \theta_{ak}^0$ dans le graphique supérieur gauche. Les estimations du graphique inférieur droit sont établies d'après les chiffres de population finale imputés, désignées par ImpGSPREE, qui sont calculés à l'étape B de l'algorithme EMQVP. Les estimations pour les municipalités les plus grandes sont semblables à celles données par SupGSPREE et la variation des changements par rapport aux proportions fondées sur le recensement est semblable à DirSPREE.

5.5 Estimation de l'EQMCP

L'EQMCP approximative des compositions ImpGSPREE peut être calculée de manière semblable à celle décrite à la section 3. Soit X_{ak}^* le dénombrement par ImpGSPREE et X_{ak}^* le meilleur prédicteur basé sur la loi conditionnelle connue de X_a sachant (y_a, m_a) . Nous avons

$$\text{EQMCP}(X_a^*) \approx E\{(X_a^* - X_a)(X_a^* - X_a)^T | y_a, m_a\} + E\{(X_a^* - X_a)(X_a^* - X_a)^T\}.$$

En outre, soit $\bar{\phi}$ l'estimation hypothétique de ϕ basée sur les données complètes $\mathbf{x} = \mathbf{X}$, et soit $\bar{\psi}$ l'estimation de ψ basée sur les données observées. Soit \bar{Q} et \bar{Q}_2 les matrices jacobiniennes des dérivées partielles $\partial X_a^* / \partial \phi$ et $\partial X_a^* / \partial \psi$, respectivement. Nous avons

$$E\{(\bar{X}_a - X_a)(\bar{X}_a - X_a)^T\}$$

$$\approx E\{(\bar{X}_a - X_a)(\bar{X}_a - X_a)^T\}$$

$$+ E\{(\bar{X}_a - X_a)(\bar{X}_a - X_a)^T | \mathbf{X}\}$$

$$\approx \bar{Q}_1 \text{Cov}(\bar{\phi}, \bar{\psi}) \bar{Q}_1^T + \bar{Q}_2 \text{Cov}(\bar{\psi}, \bar{\psi} | \mathbf{X}) \bar{Q}_2^T.$$

Six estimateurs différents de la proportion de ménages d'une seule personne (c'est-à-dire pour $k = 1$) sont illustrés à la figure 1.

Pour commencer, nous avons les proportions directes

d'après le registre θ_{1i}^0 dans le graphique situé en haut à gauche et les proportions « observées » θ_{1i}^a en haut à droite.

En moyenne, la proportion fondée sur le registre complet est

plus élevée que celle fondée sur le Recensement de 2001,

tandis que la proportion fondée sur les données

« observées » uniquement est un peu plus faible. Cela

démontre que les NII manquants sont informatifs, comme

nous l'avons expliqué plus haut. La prise en compte des

ménages enregistrés pour lesquels le NII manque accroît la

proportion de ménages d'une seule personne. Mais le

résultat n'est pas plausible dans certaines des municipalités

les plus grandes. Naturellement, un biais important existe

aussi parmi les municipalités plus petites, mais celles-ci ne

sont pas aussi faciles à déceler dans un graphique tel que

celui-ci.

Ensuite, dans le graphique situé au milieu à gauche de la

figure 1, les estimations sont obtenues par la méthode SPREE

en utilisant les chiffres de recensement $\{X_{1i}^0\}$ comme

valeurs de départ. Pour le simple tableau à double entrée que

nous avons ici, cela donne un ajustement presque constant des

proportions d'après le recensement, ainsi qu'un changement

négligeable de la variation entre domaines. Dans le graphique

du milieu à droite, les estimations sont obtenues par la

méthode SPREE en utilisant le tableau « observé » $\{v_{1i}^a\}$

comme valeurs de départ. Notons que partir des dénombre-

ments d'échantillon observés serait trop instable pour être

utile dans les situations habituelles d'échantillonnage, mais il

s'agit d'une option viable ici à cause de la grande quantité de

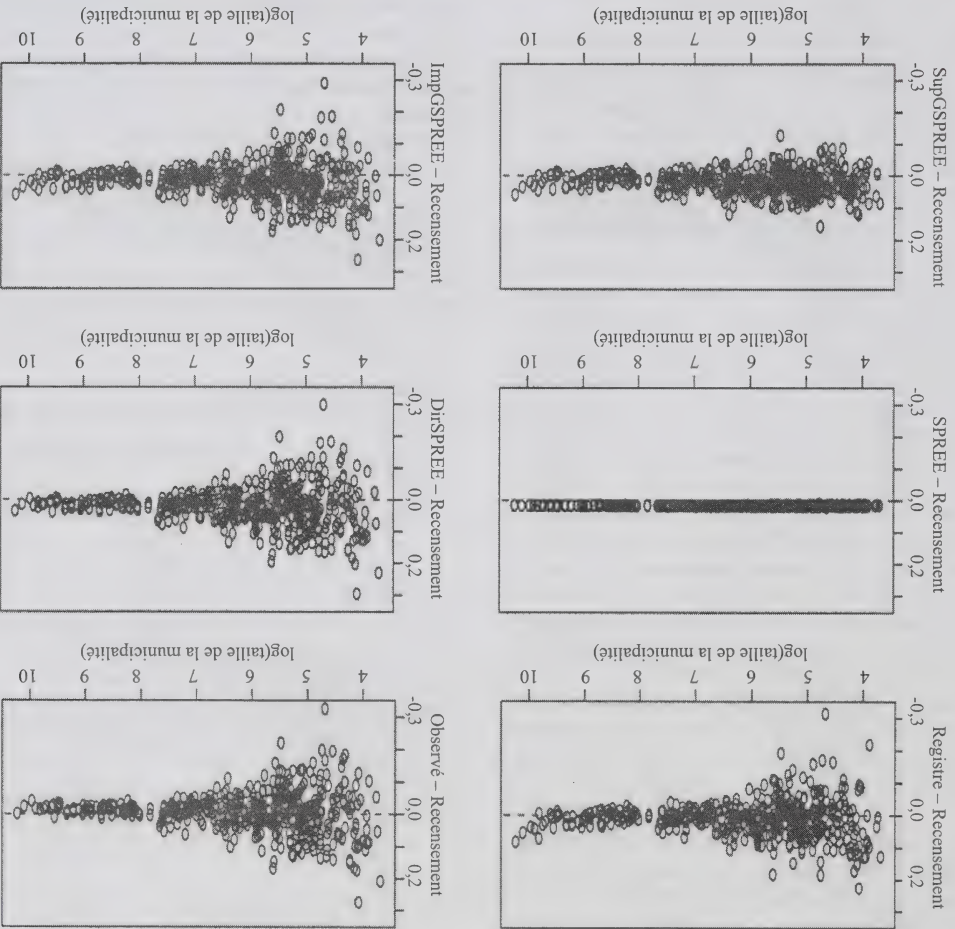
5.3 Configuration du modèle

Des diagrammes de dispersion des interactions des premier ordre $\{\alpha_{ak}^x\}$ fondées sur le registre en fonction des interactions $\{\alpha_{ak}^x\}$ fondées sur le recensement motivent l'utilisation du modèle MLMIP défini par (4). Afin de choisir entre les modèles MSMLG (2) et MMLSM (7), nous examinons la différence entre la proportion fondée sur le registre θ_{ak}^x et la proportion correspondante fondée sur le recensement θ_{ak}^0 , c'est-à-dire $\theta_{ak}^x - \theta_{ak}^0$, représentée

pour y_{ak} .

graphiquement en fonction de $\log X_a$: le cas où $k = 1$ est représenté dans le graphique supérieur gauche de la figure 1. Manifestement, la variance de la différence augmente quand X_a diminue et n'est pas constante de X_a . Notons que nous avons affaire à une estimation à un niveau d'aggrégation très faible, où par exemple la valeur médiane de tout $\{X_{ak}^x\}$ n'est que de 70. Par conséquent, nous adoptons le modèle (7) pour θ_{ak}^x , la quasi-vraisemblance (5) pour $X_{ak}^x = x_{ak}$ et la quasi-vraisemblance (8) et le modèle (9)

Figure 1 Différence entre les estimations de la proportion de ménages d'une seule personne et les chiffres de recensement en 2001 en fonction du logarithme de la taille de la municipalité : ménages enregistrés (en haut à gauche), SPREE fondée sur le recensement (milieu à droite), SupGSPREE des proportions dans la superpopulation (en bas à gauche) et ImpGSPREE des proportions en population finie imputées (en bas à droite). La droite en trait interrompu indique l'absence de différence



enfants et d'un faible taux d'enregistrement des NIL (effectivement, le plus faible dans le pays).

5.2 Configuration des données

Nous illustrerons notre approche en utilisant ces données de registre sur les ménages. La population cible contient toutes les personnes vivant à une adresse de logements multiples au début de l'année 2005 et qui n'appartiennent pas à un ménage formé de gens mariés ou de conjoints enregistrés ; ce dernier type de ménage est exclu parce qu'il n'est pas essentiel de connaître le NIL pour compiler les données sur les ménages de ces personnes. Il n'existe aucune distinction entre la population finie et l'échantillon dans ce cas, c'est-à-dire $\mathbf{X} = \mathbf{x}$. Les ménages dont le NIL est enregistré sont traités comme étant l'échantillon « observé » \mathbf{y} , tandis que ceux dont le NIL n'est pas enregistré sont considérés comme étant les observations manquantes. De cette façon, la population est constituée de 713 387 personnes, dont 558 136 possèdent un NIL enregistré. Le taux global d'observations manquantes est d'environ 22 %.

Poisons que les municipalités sont les petits domaines dans la présente étude, où $N = 433$. Les ménages sont classés en quatre catégories : $k = 1$ pour « personne seule », $k = 2$ pour « parent seul », $k = 3$ pour « cohabitants » et $k = 4$ pour « autre », d'où $K = 4$. Soit n_{ak} le nombre de ménages dans la $(a, k)^e$ cellule et soit n_{ak}^* le nombre correspondant de ménages « observés ». Notons que le nombre total de personnes est connu dans chaque domaine, mais non le nombre total de ménages. Cependant, à condition de connaître la probabilité d'enregistrement du NIL propre à la cellule, un estimateur de N_{ak} fondé sur $X_{ak} = n_{ak} X_{ak}^* / y_{ak}$. Nous nous con-

centrons donc ici sur l'estimation de X_{ak} .

Soit $\{X_{ak}^*\}$ les fréquences par cellule correspondantes tirées du dernier recensement effectué en 2001. Soit $X_{ak}^* = y_{ak} + m_{ak}$, les fréquences d'après le registre en 2005, où m_{ak} est le nombre de personnes n'ayant pas de NIL. Un ménage enregistré peut être considéré comme une forme de ménage imputé qui peut présenter une absence informative de NIL. Le total de domaine d'après le registre est correct, c'est-à-dire $X_{ak}^* = X_{ak}$, et les totaux nationaux $\{X_k^*\}$ sont considérés comme étant acceptables. La question est de savoir si les estimations de $\{X_{ak}^*\}$ peuvent être calculées en se basant sur le \mathbf{y} « observé » et la structure de répartition différentielle des NIL, $\{X_k^*\}$ et $\{X_k^*\}$, qui tient mieux compte de l'absence

5. Exemple : composition des ménages sur petits domaines fondée sur des données de registre

5.1 Données de registre sur les ménages

Les données de registre sur les ménages ont pris beaucoup d'expansion en Norvège. L'un des objectifs est de produire des statistiques détaillées sur les ménages qui ne peuvent habituellement être obtenues qu'à partir du recensement. À cette fin, l'enregistrement d'un numéro d'identification du logement a été instauré à l'occasion du dernier recensement en 2001. Les travaux ne sont pas encore achevés et le NIL manque encore pour environ 6 % des habitants du pays. Le taux de données manquantes est différentiel, car il varie selon le type de ménage, ainsi que selon la municipalité, cette dernière variation reflétant l'effort global des administrations locales en ce qui concerne l'enregistrement des NIL.

Un registre des ménages peut être compilé durant une année postcensitaire en se basant sur un certain nombre de sources de données. Les plus importantes comprennent le registre central de population (RCP), le registre des NIL et le fichier des ménages du recensement (MH01). Même sans le NIL, les données sur un ménage enregistré peuvent être compilées en se basant sur d'autres renseignements disponibles. Mais les résultats sont entachés d'un sous-enregistrement informatif du NIL. Par exemple, les cohabitants sans enfants sont une source typique de biais, parce que ces couples figurent comme des ménages composés de deux personnes seules dans le RCP, à moins qu'ils n'aient déjà été identifiés comme un ménage dans le fichier MR01. Néanmoins, des comparaisons historiques ainsi qu'à travers le pays donnent à penser que les totaux nationaux sont acceptables. Un problème plus urgent se pose aux niveaux d'agrégation plus faibles. Par exemple, les changements par rapport au Recensement de 2001 sont vraisemblablement importants dans certaines municipalités, y compris la capitale Oslo, où l'accroissement de la proportion de ménages d'une seule personne est presque trois fois plus élevé que dans le reste du pays – voir le graphique supérieur gauche de la figure 1. En outre, une grande partie du problème que pose Oslo s'explique par une combinaison d'une forte proportion de cohabitants sans

Le reste s'ensuit comme plus haut, où μ_X^a est estimé directement sous le modèle MSMG.

$$\mu_X^a = H^a \zeta + (B^a G_B^a + \tilde{Q} R_X^a \tilde{Q}^T / V^a) (\mathbf{z}_a - H^a \zeta). \quad (19)$$

où $\mathbf{e}_X^a = \tilde{Q}(\theta_X^a - \theta_X)$ et $\mathbf{e}_X^{aX} = \tilde{Q}(\mathbf{t}_X^a - \mathbf{t}_X)$, Par conséquent, nous avons $R_X^a = R_X^a + P_X^{aX}$, où $R_X^a = \text{Cov}(\theta_X^a, \theta_X^a | \theta_X)$ et $R_X^{aX} = \text{Cov}(\mathbf{t}_X^a, \mathbf{t}_X^a | \theta_X^a)$. Il s'ensuit que

covariance $\text{Cov}(\mathbf{v}_a | \mathbf{z}_a)$ ne dépend ni de \mathbf{z}_a ni de \mathbf{x}_a .

Cela est commode, parce que nous avons alors

$$h_a(\mathbf{x}_a; \zeta, \delta) \approx B_a^a \text{Cov}(\mathbf{v}_a | \mathbf{z}_a) B_a^a \quad (16)$$

où $V_a = B_a^a G B_a^a + \bar{Q} R_a \bar{Q}^T$ est la matrice de covariance marginale de \mathbf{z}_a .

Ensuite, prenons h_a donné par (13). Soit $\phi = (\zeta^T, \delta^T)^T$. Le développement de $\hat{\phi}$ autour de ϕ donne $\hat{\mu}_a - \mu_a \approx \hat{\mu}_a'(\phi - \phi)$, où $\hat{\mu}_a' = \partial \hat{\mu}_a / \partial \phi$, tel que

$$h_{z_a} \approx \hat{\mu}_a' \text{Cov}(\hat{\phi}, \hat{\phi}) \hat{\mu}_a'^T. \quad (17)$$

En partant de (6), nous dérivons $\hat{\mu}_a = H_a^a \zeta + D_a^a \mathbf{u}_a$, où $D_a^a = B_a^a G B_a^a V_a^{-1}$ et $\mathbf{u}_a = \mathbf{z}_a - H_a^a \zeta$. Désignons par I la matrice identité. Dans $\hat{\mu}_a'$, les dérivées partielles sont

$$\partial \hat{\mu}_a / \partial \zeta = (I - D_a^a) H_a^a$$

et

$$\partial \hat{\mu}_a / \partial \delta_j = (\partial D_a^a / \partial \delta_j) \mathbf{u}_a = (I - D_a^a) B_a^a (\partial G / \partial \delta_j) B_a^a V_a^{-1} \mathbf{u}_a$$

où δ_j est le j^{e} paramètre de variance dans la matrice de covariance $G(\delta)$ de \mathbf{v}_a . Pour obtenir $\text{Cov}(\hat{\phi}, \hat{\phi})$, supposons que l'approche QVP est basée sur la quasi-vraisemblance qui suit

$$\ell = \sum_a \ell_a$$

et

$$\ell_a = -\frac{1}{2} \log |V_a| - \frac{1}{2} (\mathbf{z}_a - H_a^a \zeta)^T V_a^{-1} (\mathbf{z}_a - H_a^a \zeta).$$

La formule dite sandwich donne alors

$$\text{Cov}(\hat{\phi}, \hat{\phi}) = \left(-\frac{\partial^2 \ell}{\partial^2 \hat{\phi}} \right)^{-1} \left\{ \sum_a \left(\frac{\partial \ell}{\partial \hat{\phi}} \right)_a \right\} \left(\frac{\partial \ell}{\partial \hat{\phi}} \right)_a^T \left(-\frac{\partial^2 \ell}{\partial^2 \hat{\phi}} \right)^{-1}.$$

Enfin, prenons h_{z_a} donné par (14). Comme ci-dessus, nous avons $\hat{\mu}_a = (I - D_a^a) H_a^a \zeta + D_a^a \mathbf{z}_a$ évalué à $\phi = \hat{\phi}$, et

pour $\hat{\mathbf{x}}_a = E(\mathbf{x}_a | \mathbf{y}_a, m_a; \hat{\phi}, \hat{\psi})$. En développant $\hat{\phi}$ autour de $\hat{\phi}$ et en ne gardant que le premier terme, nous obtenons

$$\hat{\mu}_a - \mu_a \approx \hat{\mu}_a' - \hat{\mu}_a = D_a^a (\mathbf{z}_a - \mathbf{z}_a)$$

où $\hat{\mu}_a' = (I - D_a^a) H_a^a \zeta + D_a^a \mathbf{z}_a$, et \mathbf{z}_a est dérivé de $\hat{\mathbf{t}}_a = \hat{\mathbf{t}}(\hat{\mathbf{x}}_a)$ pour $\hat{\mathbf{x}}_a = E(\mathbf{x}_a | \mathbf{y}_a, m_a; \hat{\phi}, \hat{\psi})$. Autrement dit, nous ignorons les termes comportant $\hat{\phi} - \phi$. La variation persistante dans \mathbf{z}_a est due à l'estimation du modèle avec données manquantes uniquement. En développant $\hat{\psi}$ autour de ψ , nous obtenons, en appliquant la règle de la chaîne,

$$\begin{aligned} \mathbf{z}_a &= H_a^a \zeta + B_a^a \mathbf{v}_a + \mathbf{e}_a = H_a^a \zeta + B_a^a \mathbf{v}_a + \mathbf{e}_a^a + \mathbf{e}_a^{a|x} \\ &= \mu_a^a \zeta + \mathbf{v}_a^a + \mathbf{e}_a^{a|x} \end{aligned}$$

Supposons ensuite que nous avons estimé le modèle MMLSM défini par (7) et combiné à (5). Nous pouvons exprimer l'estimation d'intérêt, c'est-à-dire $\{\mu_a^a\}$, en fonction de \mathbf{z}_a défini comme étant

estimations de $\{\theta_a^a\}$ présentant un biais plus faible et, par conséquent, des biais plus faible produisent des estimations de $\{\alpha_a^a\}$ attendons à ce que des estimations de $\{X_a^a\}$ présentant un estimation des effets principaux $\{\alpha_a^a\}$. Donc, nous l'estimation finale θ_a^a est due à la différence entre les différences entre l'estimation directe du modèle $\hat{\theta}_a^a$ et $\alpha_a^a = \hat{\alpha}_a^a$. En vertu de l'identité log-linéaire (3), la estimées $\hat{\alpha}_a^a$ sont préservées dans l'API, c'est-à-dire agrégées $\sum_a X_a^a \theta_a^a$. Cela tient au fait que les interactions moins biaisées que les estimations sur petits domaines totaux marginaux estimés soient jugés plus fiables et/ou encore d'être pris en considération à condition que ces approchée pour le niveau d'aggrégation. L'API mérite d'enquête disponibles, séparément en utilisant une méthode totaux marginaux $\{X_a^a\}$ en se fondant sur les données totaux de domaine $\{X_a^a\}$, mais que l'on doit estimer les En pratique, il arrive souvent que l'on connaisse les valeurs de départ pour l'API.

des estimations directes du modèle $\hat{\theta}_a^a$ qui ont fourni les pondantes, désignées par $\theta_a^a = X_a^a / \sum_j X_j^a$, qui diffèrent désignées par $\hat{\mathbf{X}} = \{X_a^a\}$, et les compositions correspondantes les dénominateurs estimés sur petits domaines, interactions ont été estimées. À la convergence, nous à SPREB, qui débute par le tableau auxiliaire \mathbf{X}_0^a , est que les en partant du tableau estimé $\{\theta_a^a\}$. La différence par rapport logique d'appliquer l'ajustement proportionnel itératif (API) Si les totaux marginaux X_a^a et X_j^a sont connus, il est donné $\hat{\mu}_a^a = \exp(\hat{\mu}_a^a / \sum_j \exp(\hat{\mu}_j^a))$.

MMLSG défini par (2) et combiné à (5), ce qui nous a Supposons d'abord que nous avons estimé le modèle

4.3 Estimation de la composition sur petits domaines

$\text{Cov}(\hat{\phi}, \hat{\phi})$ susmentionné.

le modèle conditionnel de \mathbf{y} sachant \mathbf{x} , similairement à Tandis que la formule sandwich donne $\text{Cov}(\hat{\psi}, \hat{\psi} | \mathbf{x})$ sous où nous supposons que $E(\hat{\psi} | \mathbf{x}) = \psi$ et $E[\mathbf{z}_a | \mathbf{x}] = \mathbf{z}_a$.

$$C_a = \left\{ D_a^a \left(\frac{\partial \alpha_a^a}{\partial \zeta} \right) \left(\frac{\partial \alpha_a^a}{\partial \zeta} \right)^T \left(\frac{\partial \alpha_a^a}{\partial \delta_j} \right) \left(\frac{\partial \alpha_a^a}{\partial \delta_j} \right)^T \left(\frac{\partial \psi}{\partial \psi} \right) \right\}_{\phi = \hat{\phi}}$$

et

$$h_{z_a} \approx C_a^a \text{Cov}(\hat{\psi}, \hat{\psi} | \mathbf{x}) C_a^a$$

quasi-vraisemblance pénalisée (MQVP). Les itérations entre les deux donnent un algorithme EMQVP.

Pour l'étape E, soit $I_{i,ak} = 1$ si l'unité échantillonnée i appartient à la (a, k) ° cellule et $I_{i,ak} = 0$ autrement. La valeur est observée à condition que $r_{i,ak} = 1$, mais est inconnue si $r_{i,ak} = 0$. Soit θ_{ak} les compositions génétiques, qui dépendent du modèle adopté. Supposons que

$$P[I_{i,ak} = 1 | i \in s] = d_{ak}\theta_{ak}$$

où s désigne l'échantillon complet et d_{ak} est une constante connue qui tient compte de l'effet de plan d'échantillonnage éventuel. Par exemple, l'échantillonnage aléatoire simple implique que $d_{ak} = 1$ pour toute (a, k) . Un exemple de situation où $d_{ak} \neq 1$ est celle où les unités d'échantillonnage sont les ménages, qui sont sélectionnés avec probabilité proportionnelle à la taille du ménage. Soit $m_{ak} = x_{ak} - Y_{ak} = \sum_{i: r_{i,ak}=0} I_{i,ak} x_{i,ak} + E(m_{ak} | m_a)$, où

$$E(m_{ak} | m_a) = \sum_{i: r_{i,ak}=0} E(I_{i,ak} | r_{i,ak} = 0) x_{i,ak}$$

$$= m_{ak} P[I_{i,ak} = 1 | r_{i,ak} = 0]$$

$$= m_a (1 - d_{ak}) d_{ak} \theta_{ak} / \left(\sum_i (1 - d_{ij}) \theta_{ij} \right). \quad (15)$$

Ayant donc « complet » les données d'échantillon, nous passons à l'étape MQVP, où nous appliquons l'algorithme MCP décrit à la section 2.1.3 au modèle avec données complètes et au modèle avec données manquantes conditionnellement aux données complètes, respectivement.

4.2 Estimation de l'EQMCP

L'évaluation de l'EQMCP aux valeurs estimées des paramètres donne une estimation par insertion des données (plug-in) de l'EQMCP. Des trois termes h , h_a est d'ordre $O_p(1)$, tandis que h_a et h_a sont d'ordre $O_p(A^{-1})$, quand le nombre de domaines tend vers l'infini mais que les tailles d'échantillon à l'intérieur du domaine demeurent bornées. Les résultats de Booth et Hobert (1998) et de Prasad et Rao (1999), obtenus dans le cas de données complètes univariées, donnent à penser que le biais dans l'estimation par insertion de données h_a est du même ordre que dans h_a et h_a . Ces auteurs ont élaboré une correction de deuxième ordre par développement en série de Taylor. Nous ne poursuivons pas l'obtention de ces asymptotiques de approximations des termes h qui accompagnent l'algorithme EMQVP sont données ci-après.

Pretons d'abord h_a donné par l'équation (12). En se basant sur les données linéarisées (6), la matrice de

De cette façon, nous obtenons une EQMCP approximative décomposée en trois parties

$$EQMCP_a \approx h_a(x_a; \zeta, \delta) + h_2(\zeta, \delta) + h_3(x_a; \zeta, \delta, \psi)$$

où ψ contient les paramètres de la loi conditionnelle de y_a sachant x_a , et

$$h_a(x_a; \zeta, \delta) = B_a \text{Cov}(v_a, v_a | x_a) B_a^T \quad \text{dét.} \quad (12)$$

$$h_2(\zeta, \delta) = E\{(\bar{u}_a - \bar{u}_a)(\bar{u}_a - \bar{u}_a)^T\} \quad \text{dét.} \quad (13)$$

$$h_3(x_a; \zeta, \delta, \psi) = E\{(\bar{u}_a - \bar{u}_a)(\bar{u}_a - \bar{u}_a)^T | x_a\}. \quad (14)$$

Les trois termes h correspondent, respectivement, à une variance de prédiction conditionnelle due aux effets aléatoires, une correction positive qui tient compte de l'incertitude dans l'estimation des paramètres basée sur les données complètes latentes, c'est-à-dire la variation d'échantillonnage, et une autre correction positive pour la variation supplémentaire due au caractère aléatoire des données manquantes. D'autres approximations sont possibles. Nous pourrions utiliser $E\{B_a \text{Cov}(v_a, v_a | x_a) B_a^T | y_a\}$ au lieu de h_a , ou remplacer h_3 par l'espérance non conditionnelle $E\{(\bar{u}_a - \bar{u}_a)(\bar{u}_a - \bar{u}_a)^T\}$. Les expressions (12) à (14) sont choisies parce qu'elles produisent une séparation nette entre la variation d'échantillonnage dans les données complètes et la variation supplémentaire due à la présence de données manquantes sachant les données complètes. La différence par rapport à l'EQMCP dans le cas des données complètes (Booth et Hobert 1998) tient au troisième terme h_3 .

4. Estimation

4.1 Estimation des paramètres

La structure des données évoque une procédure itérative semblable à l'algorithme EM (Dempster, Laird et Rubin 1977). Sachant les valeurs courantes des paramètres et les effets aléatoires, nous calculons à l'étape E le tableau à double entrée prévu conditionnel $E(x | y, m)$. À l'étape M, nous estimons les deux modèles mixtes à effets aléatoires séparément par une méthode du maximum de

Les interactions de premier ordre de $\{d^a_k\}$ sont alors données par $\alpha^a_k = -\gamma^a_k - \gamma^a_k - \gamma^a_k - \gamma^a_k$, pour les moyennes de ligne et de colonne $\bar{\gamma}^a_k$ et $\bar{\gamma}^a_k$, ainsi que pour la moyenne globale $\bar{\gamma}^a$. Ces interactions ne sont pas nulles, à moins que $\xi^a_k = \xi^a$. En vertu de (8), les interactions du tableau observé prévu sont données par

$$\alpha^a_{E(y|x,b)} = \alpha^a_x + \alpha^a_p - \alpha^a_k - \tilde{\gamma}^a_k \neq \alpha^a_x$$

de sorte que les estimations de $\{\alpha^a_x\}$ comporteront un biais si y est traité comme étant x .

Il convient de souligner qu'en ce qui concerne l'estimation des interactions, il est en principe possible de traiter le tableau observé y comme s'il s'agissait du tableau complet x sous un modèle de données manquantes particulier donné par

$$(10) \quad \log p^a_k = \xi^a_k + b^a_k.$$

Il en est ainsi parce que les interactions de premier ordre de $\{d^a_k\}$ sont toutes nulles sous (10), auquel cas nous avons $\alpha^a_{E(y|x)} = \alpha^a_x$. Si nous ne tenons pas compte des contraintes sur x^a , donné par $\mu^a = E(\mu^a | x^a, \zeta^a, \delta^a) = H^a \zeta^a + B^a E(v^a | x^a, \zeta^a, \delta^a)$, quand les paramètres sont connus. Nous avons

$$\begin{aligned} \text{EQMCP}^a = E\{E((\mu^a - \mu^a)(\mu^a - \mu^a)^T | y^a) \\ + E\{B^a \text{Cov}(v^a, v^a | x^a) B^{aT} | y^a\} \\ + E\{(\mu^a - \mu^a)(\mu^a - \mu^a)^T | y^a\} \end{aligned}$$

Pour commencer, nous introduisons une décomposition par la voie du meilleur prédicteur (MP) hypothétique fondée sur x^a , où nous ne tenons pas compte des contraintes sur x^a , donné par $\mu^a = E(\mu^a | x^a, \zeta^a, \delta^a) = H^a \zeta^a + B^a E(v^a | x^a, \zeta^a, \delta^a)$, quand les paramètres sont connus. Nous

$$\begin{aligned} E\{(\mu^a - \mu^a)(\mu^a - \mu^a)^T | y^a\} \approx E\{(\mu^a - \mu^a)(\mu^a - \mu^a)^T | x^a\} \\ \approx E\{(\mu^a - \mu^a)(\mu^a - \mu^a)^T | y^a\} \approx E\{(\mu^a - \mu^a)(\mu^a - \mu^a)^T | x^a\} \end{aligned}$$

parce que les termes $\mu^a - \mu^a$ et $\mu^a - \mu^a$ sont conditionnellement indépendants l'un de l'autre sachant x^a : $\mu^a - \mu^a$ dépend des effets aléatoires v^a , tandis que $\mu^a - \mu^a$ dépend des variations aléatoires dans les autres domaines. Ensuite, pour le deuxième terme du deuxième membre, nous introduisons une décomposition par la voie du meilleur prédicteur estimé (MP-E) hypothétique basée sur les données complètes x , désigné par $\mu^a = H^a \zeta^a + B^a v^a$, où (ζ^a, δ^a) sont les estimations des paramètres fondées sur x , et $v^a = E(v^a | x^a, \zeta^a, \delta^a)$. Nous avons

L'échelle du prédicteur linéaire. Sous forme vectorielle, les μ^a donnés par (1) appartiennent à la classe de fonctions linéaires

$$(11) \quad \mu^a = H^a \zeta^a + B^a v^a$$

où μ^a est le vecteur propre au domaine de prédicteurs linéaires, ζ^a est le vecteur des effets fixes, v^a est le vecteur des effets aléatoires propres au domaine, et H^a et B^a sont les matrices de plan correspondantes. Toutes les quantités ont été spécifiées dans (6) pour le MSMIG (2), où nous avons effectivement $B^a = B$. Nous adopterons toutefois la formulation un peu plus générale (11) dans la suite. Soit ζ^a et v^a les estimations de ζ^a et v^a , respectivement, fondées sur des observations pouvant présenter des données manquantes, désignées par y^a pour $a = 1, \dots, A$. L'EQMCP de

3. Erreurs quadratiques moyennes conditionnelles de prédiction

Enfin, nous constatons que l'introduction d'effets aléatoires relatifs aux composantes dans le modèle (9) peut causer des problèmes d'identification. Par exemple, supposons un échantillonnage aléatoire simple à partir de la population finie, auquel cas les interactions du tableau complet prévu sont données par $\alpha^a_{E(x|x)} = \alpha^a_x$. Avec l'effet aléatoire concernant les composantes b^a_k dans le modèle (9), nous avons $\log p^a_k = \xi^a_k + b^a_k + \gamma^a_k$, où $\gamma^a_k = \log(1 + \exp(\xi^a_k + b^a_k))$. Il découle de (4) et (8) que les interactions du tableau prévu $E(y | x, b)$ sont données par $\beta^a_{0^a} + \gamma^a_k + b^a_k - \gamma^a_k$. Mais il n'existe aucune information dans les données observées permettant de faire la distinction entre les deux effets aléatoires v^a_k et b^a_k .

Nous adoptons l'approche de Booth et Hobert (1998) et utilisons l'EQMCP comme mesure de l'incertitude de la prédiction. Comme eux, nous considérons l'EQMCP sur

La première approximation est correcte jusqu'à l'ordre $O_p(F^{-1})$ et peut être justifiée à mesure que le nombre de

$(X_{a1}^T, \dots, X_{ak}^T)$ suivent la loi multinomiale avec les paramètres $(\theta_1^{ak}, \dots, \theta_K^{ak})^T$. Un modèle mixte log-standardisé multinomial (MMLSM) de $\{\theta^{ak}\}$ est donné par

$$\mu^{ak} = \lambda_k + \beta \mu_0^{ak} + \nu^{ak} \quad (7)$$
$$\sum_K \lambda_k = 0 \quad \text{et} \quad \sum_K \nu^{ak} = 0$$

où μ^{ak} est donné par θ_a par la voie de la fonction lien logsm.

Contrairement à l'équation (2) du MSMG, l'équation (7) définit un modèle de régression. Dans la situation d'enquête par sondage, nous avons alors le choix entre trois hiérarchies distinctes :

1. Émettre l'hypothèse du MSMG (2) pour la population finie et du modèle de quasi-vraisemblance (5) pour l'échantillon, ce qui donne l'approche GSPRE de Zhang et Chambers (2004).

2. Émettre l'hypothèse du MMLSM (7) pour la superpopulation et modéliser les données d'échantillon t_a en se fondant directement sur θ_a , ce qui donne une approche à deux niveaux fondée purement sur un modèle.

3. Émettre l'hypothèse du MMLSM (7) pour la superpopulation, supposer que les totaux de population finie X_a suivent la loi multinomiale sachant θ_a , et émettre l'hypothèse du modèle de quasi-vraisemblance (5) sachant X_a , ce qui donne un modèle général à trois niveaux.

En pratique, à condition que la population finie soit grande, la différence est faible si l'on adopte l'approche GSPRE, ce qui évite de devoir traiter explicitement un niveau hiérarchique supplémentaire. Cependant, la distinction entre (2) et (7) devient nécessaire si les domaines sont si petits que la variation stochastique dans X_a n'est pas négligeable comparativement à la variation d'échantillonnage dans x_a (ou t_a). Dans l'application que nous décrivons plus loin, nous utilisons des données de registre qui nous auraient donné les dénombrements de population d'intérêt $\{X_a^{ak}\}$ si l'y avait pas eu de données manquantes. En outre, le niveau d'aggrégation sur petit domaine est si fin que la variation stochastique dans X_a ne peut pas être ignorée. Par conséquent, nous adoptons l'approche GSPRE a) en choisissant le modèle MMLSM (7) au lieu du modèle MSMG (2) et b) en modélisant X_a comme un « échantillon », quoique de très grande taille, directement à partir de la superpopulation.

2.2 Un modèle mixte à effets aléatoires des données manquantes

Les données manquantes ajoutent un niveau de variation stochastique en sus de celle des données complètes sous-jacentes. Dans l'exposé qui suit, considérons que les dénombrements d'échantillons $\{x_a^{ak}\}$ représentent les données complètes, ce qui est la situation la plus fréquente en pratique. L'application que nous décrivons à la section 5 peut être considérée comme un cas particulier où $X = x$. Désignons par $y_a = (y_{a1}, \dots, y_{ak})^T$ les fréquences observées par cellule pour $a = 1, \dots, A$. Supposons que, conditionnellement à x_a , et à un effet aléatoire b_a ,

$$E(y_{ak} | x_a^{ak}, b_a) = x_a^{ak} p^{ak} \quad (8)$$
$$V(y_{ak} | x_a^{ak}, b_a) = \nu^2 c^{ak} p^{ak} (1 - p^{ak})$$

où c^{ak} est une constante connue et ν^2 est le paramètre de dispersion. Nous supposons que y_a^{ak} est indépendant de y_{aj} pour $k \neq j$, c'est-à-dire que les données manquantes sont indépendantes d'une cellule à l'autre. Donnons aux unités comprises dans la cellule d'échantillon complète (a, k) l'indice $i = 1, \dots, m^{ak}$. Soit $r_{i,ak} = 1$ si la i^e unité est observée et $r_{i,ak} = 0$ si elle est manquante. Le paramètre p^{ak} est la probabilité hypothétique que $r_{i,ak} = 1$ à l'intérieur de la cellule (a, k) . Pour le montrer, posons que $x_{i,ak}$ est la contribution de la i^e unité à x_a^{ak} , c'est-à-dire que

$$x_a^{ak} = \sum_{i=1}^{m^{ak}} x_{i,ak}, \text{ de sorte que } y_{ak} = \sum_{i=1}^{m^{ak}} r_{i,ak} x_{i,ak} \text{ et}$$
$$E(y_{ak} | x_{i,ak}, \dots, x_{m^{ak},ak}, b_a) = \sum_{i=1}^{m^{ak}} x_{i,ak} E(r_{i,ak} | b_a) = \sum_{i=1}^{m^{ak}} x_{i,ak} p^{ak} = x_a^{ak} p^{ak}.$$

Soulignons que p^{ak} ne dépend pas de la valeur de $x_{i,ak}$, mais uniquement de la position de l'unité dans le tableau à double entrée. Nous supposons que p^{ak} dépend de b_a par la voie de la fonction de lien logarithmique donné par

$$\eta^{ak} = \log(p^{ak} / (1 - p^{ak})) = \zeta_k + b_a \quad (9)$$

Les effets fixes ζ_k permettent que la probabilité que des observations manquent dépende des catégories d'intérêt et l'effet aléatoire au niveau du domaine b_a permet en outre qu'elle varie d'un domaine à l'autre. Manifestement, sous les hypothèses (8) et (9), les données manquantes causent un biais dans les estimations des λ_k si le tableau observé y est traité comme s'il était complet. De surcroît, cela fausse l'estimation des interactions de premier ordre $\{\alpha_X^{ak}\}$. Nous avons

$$\log p^{ak} = (\zeta_k + b_a) - \gamma^{ak} \text{ ou } \gamma^{ak} = \log(1 + \exp(\zeta_k + b_a)).$$

Une interprétation importante du modèle (2) en ce qui concerne les interactions log-linéaires de $\{\theta_{ak}\}$ résulte du choix de la fonction de lien (1), c'est-à-dire

$$\mu_X^{ak} = \alpha_k + \alpha_X^{ak} \quad (3)$$

où, en vertu de la théorie standard des modèles log-linéaires (par exemple Agresti 2002), nous avons

$$\log X^{ak} = \log X_a + \log \theta_X^{ak} = \alpha_X^0 + \alpha_X^a + \alpha_X^k + \alpha_X^{ak}$$

pour $\alpha_X^0 = (AK)^{-1} \sum_a \sum_k \log X^{ak}$, et $\alpha_X^a = A^{-1} \sum_k \log X^{ak} - \alpha_X^0$, et $\alpha_X^k = A^{-1} \sum_a \log X^{ak} - \alpha_X^0$, tels que $\sum_a \alpha_X^a = \sum_k \alpha_X^k = \sum_a \alpha_X^{ak} = \sum_k \alpha_X^{ak} = 0$. Nous appelons (3) l'identité log-linéaire et nous désignons les paramètres log-linéaires α_X^{ak} comme étant les interactions (de premier ordre) des compositions θ_X^{ak} ainsi que des dénombrements X^{ak} . Une identité semblable est vérifiée pour μ_a^{ak} . Zhang et Chambers (2004) ont montré que le MSMLG est équivalent au modèle mixte à interactions proportionnelles (MIMP) suivant

$$\alpha_X^{ak} = \beta \alpha_a^k + \nu_a^k + O^p(A^{-1/2}). \quad (4)$$

Dans (2), les paramètres λ_k ne comportent aucune restriction de modélisation outre le MMIP, et ils n'affectent pas les interactions. Le paramètre β est appelé coefficient de proportionnalité. Clairement, l'approche SPREE fondée directement sur la structure des associations $\{X^{ak}\}$ revient à poser que $\beta \equiv 1$ et $\nu_a^k \equiv 0$. Par conséquent, nous donnons au modèle (2) le nom de modèle GSPREE, qui contient les extensions à effets fixes et aléatoires du modèle SPREE.

2.1.2 Modèle pour échantillon

Pour établir la spécification du modèle, nous supposons que nous avons les classifications d'échantillon $\mathbf{x} = \{x_{ak}\}$. Soit

$$\mathbf{t}_a = (t_{a1}, \dots, t_{ak}, \dots, t_{aK})^T = (t_1(\mathbf{x}), \dots, t_K(\mathbf{x}))^T$$

tel que $E(t_{ak} | \mathbf{v}) = E(t_{ak} | \mathbf{X}) = \theta_X^{ak}$, où $\mathbf{v} = \{\nu_a^k\}$. L'espérance est habituellement calculée par rapport au plan d'échantillonnage. Cependant, elle peut aussi être calculée sous un modèle approprié de la distribution d'échantillonnage, comme un modèle multinomial pour \mathbf{x}_a sous la contrainte d'un échantillonnage aléatoire simple dans chaque domaine. Par conséquent, nous ne faisons aucune distinction dans la notation.

Nous supposons que \mathbf{t}_a est indépendant de $\mathbf{t}_{a'}$ pour $a \neq a'$, et posons que

$$I(t_{ak}) = \nu_a^k \omega_k(\mathbf{X}_a) \quad \text{et} \quad \text{Cov}(t_{ak}, t_{a'k'}) = \nu_a^k \omega_{kk'}(\mathbf{X}_a) \quad (5)$$

où $\omega_k(\cdot)$ et $\omega_{kk'}(\cdot)$ sont les fonctions de variance et de covariance spécifiées, et ν_a^k est le paramètre de dispersion

qui peut être connu ou non. Il s'agit essentiellement des conditions de quasi-vraisemblance pour des données dépendantes (McCullagh et Nelder 1989). La dépendance à l'égard de \mathbf{X}_a nous permet d'intégrer l'effet du plan d'échantillonnage, auquel cas, dans (5), les espérances peuvent être évaluées par rapport à la distribution d'échantillonnage. Il s'agit d'une raison importante pour laquelle nous ne supposons pas directement que la distribution de \mathbf{t}_a appartient à la famille exponentielle, comme cela est le cas par exemple dans les modèles linéaires mixtes généralisés (Breslow et Clayton 1993).

2.1.3 Estimation des paramètres

Zhang et Chambers (2004) décrivent pour le MSMLG donné par (2) un algorithme basé sur les moindres carrés pondérés itératifs (MCPI) qui est une variation de l'approche de quasi-vraisemblance pénalisée (QVP) (Schall 1991; Breslow et Clayton 1993). Soit $\mu_a = (\mu_a^1, \dots, \mu_a^K)^T$. Le MSMLG (2) peut être donné formellement sous la forme

$$\mu_a = g(\theta_a) = H_a \zeta + B \nu_a^{(1)}$$

où $g(\theta_a)$ est la fonction de lien logsm, $\zeta = (\lambda_2, \dots, \lambda_K, \beta)^T$ et $\nu_a^{(1)} = (\nu_a^{21}, \dots, \nu_a^{K1})^T$. La matrice de plan H_a de dimensions $K \times K$ et la matrice de plan B de dimensions $K \times (K-1)$ sont, respectivement,

$$H_a = [B_{K \times K-1} \quad \mu_a^0] \quad \text{et} \quad B = \begin{pmatrix} -1 & \dots & -1 \\ I_{K-1 \times K-1} \end{pmatrix},$$

où $\mathbf{1}$ est un vecteur de 1 et I est une matrice identité. Définissons les variables de travail

$$\mathbf{z}_a = \mu_a + \mathbf{e}_a = H_a \zeta + B \nu_a + \mathbf{e}_a \quad \text{et} \quad \mathbf{e}_a = Q(\mathbf{t}_a - \theta_X^a) \quad (6)$$

où $\bar{Q} = \partial \mu_a^a / \partial \theta_X^a$ est la matrice jacobienne des dérivées partielles. Désignons par R_a la matrice de covariance conditionnelle de \mathbf{t}_a sachant θ_X^a défini par (5). Sous l'approche QVP, nous supposons que \mathbf{e}_a suit une loi normale multivariée approximative avec matrice de covariance $\bar{Q} R_a \bar{Q}^T$, et nous appliquons les méthodes standard pour les modèles mixtes linéaires (ML) aux données linéarisées (6). Les variances de l'approche QVP diffèrent en ce qui concerne l'estimation des paramètres de variance δ . Nous omettons les détails ici.

2.1.4 De la hiérarchie des modèles

Le MSMLG (2) est spécifié au niveau de la population finie. De manière plus générale, nous pouvons considérer que la population finie $\{X^{ak}\}$ est produite aléatoirement à partir d'une superpopulation infinie. Soit θ_a^k la probabilité intra-domaine qu'une unité de la superpopulation appartienne à la cellule (a, k) , où $\sum_k \theta_a^k = 1$. Conditionnellement à

2. Modélisation mixte double

2.1 Modèle mixte à effets aléatoires dans le cas de

données complètes

2.1.1 Modèles pour population finie

Les dénombrements sur petits domaines peuvent être disposés dans un tableau de contingence à double entrée désigné par $\mathbf{X} = \{X_{ak}\}$, où $a = 1, \dots, A$ sont les indices des petits domaines et $k = 1, \dots, K$, ceux des catégories d'intérêt. Les grandeurs que nous souhaitons estimer sont les proportions intra-domaine données par

$$\theta_X^{ak} = X_{ak} / X_{a\cdot} = X_{ak} / \sum_{k=1}^K X_{ak}$$

que nous appelons compositions, puisque $\sum_k \theta_X^{ak} = 1$. Habituellement, sous l'approche GSPREE, nous supposons que les totaux marginaux $\{X_{a\cdot}\}$ et $\{X_{\cdot k}\}$, également appelés structure de répartition, sont connus ou peuvent être estimés fiablement, auquel cas l'estimation $\{\theta_X^{ak}\}$ est équivalente à l'estimation $\{X_{ak}\}$. Pour simplifier, nous ne faisons alors aucune distinction entre les dénombrements et les compositions dans l'exposé. Sinon, sans la structure de répartition, nous pouvons encore utiliser notre approche pour estimer $\{\theta_X^{ak}\}$, mais non $\{X_{ak}\}$.

Supposons que nous disposons d'un tableau auxiliaire de la même variable, désigné par $\mathbf{X}^0 = \{X_{ak}^0\}$ et des proportions intra-domaine correspondantes $\{\theta_{X^0}^{ak}\}$. Pour modéliser $\theta_X^{ak} = (\theta_X^{a1}, \dots, \theta_X^{aK})^T$, nous utilisons la fonction

de lien *log-standardisée multinomiale* (*logsm*) donnée par

$$(1) \quad \mu_X^{ak} = \log \theta_X^{ak} = \log \theta_{X^0}^{ak} - K^{-1} \sum_{j=1}^J \log \theta_X^{aj}$$

et faisons de même pour $\mu_{X^0}^{ak}$ et $\theta_{X^0}^{ak}$. Zhang et Chambers (2004) ont introduit le modèle structurel mixte linéaire généralisé (MSMLG)

$$\mu_X^{ak} = \gamma_k + \beta \mu_{X^0}^{ak} + v_{ak}$$

où

$$\sum_K \gamma_k = 0 \quad \text{et} \quad \sum_K v_{ak} = 0$$

et $v_{a(1)} = (v_{a2}, \dots, v_{aK})^T$ suit une loi normale multivariée avec matrice de covariance $G = G(\delta)$, où δ contient les paramètres de variance. Notons que (2) ne contient aucun terme partiellicier au domaine, parce que $\sum_k \mu_{X^0}^{ak} = \sum_k \mu_{X^0}^{ak} = 0$. Le terme « structurel » fait référence au fait qu'il s'agit d'un modèle direct des paramètres de population finie $\{\theta_X^{ak}\}$, bien qu'il ne soit pas courant d'insister sur ce fait dans la littérature traitant de l'estimation sur petits domaines. Par exemple, le modèle bien connu de Fay-Herriot (Fay et Herriot 1979) est « structurel » dans ce sens.

domaines, et ce pour des variables qui ne correspondent pas nécessairement à celles d'intérêt pour les petits domaines. S'ils sont disponibles, les totaux ajustés peuvent être intégrés dans l'approche GSPREE comme totaux marginaux pour l'ajustement proportionnel itératif (API). Cependant, la modélisation des probabilités différentielles que des données manquent dans les divers petits domaines demeure généralement intéressante.

Il convient aussi de souligner qu'en tant que telles, les données manquantes informatives rendent plus difficile l'évaluation du biais éventuel de toute approche d'estimation. L'approche SPREE peut comporter un biais de deux facteurs : i) les hypothèses de modèle log-linéaire restent sous-jacentes ne soit vraisemblablement pas réalistes, ii) l'API direct pourrait ne pas tenir compte adéquatement des probabilités différentielles d'existence de données manquantes. L'approche de modélisation mixte double proposée résout le problème (i) par modélisation GSPREE des données complètes sous-jacentes et s'occupe du problème (ii) par introduction d'un modèle de création de données manquantes plus souple, dont nous discuterons à la section 2.2. Néanmoins, un certain biais persistera vraisemblablement. Puisque l'estimation des paramètres du modèle et des effets aléatoires est plus compliquée sous l'approche de modélisation mixte double, d'autres méthodes d'estimation permettant de préserver la simplicité de calcul de SPREE tout en faisant une correction plus appropriée pour les données manquantes informatives mériteraient d'être étudiées dans l'avenir.

En ce qui concerne l'évaluation de l'incertitude de l'estimation, Booth et Hobert (1998) ont défendu l'utilisation de l'erreur quadratique moyenne conditionnelle de prédiction (EQMCP) sachant les données observées. Nous étendons leur approche et calculons une EQMCP approximative dans la situation courante de données incomplètes multivariées. Nous obtenons ainsi une décomposition de l'EQMCP en trois parties, qui correspondent à une variance de prédiction *naïve*, une correction positive qui tient compte de l'incertitude hypothétique d'estimation des paramètres fondée sur les données complètes latentes et une autre correction positive pour la variance supplémentaire due aux données manquantes. Les détails sont donnés à la section 3.

Les méthodes d'estimation des paramètres, de l'EQMCP et des compositions sur petits domaines sont décrites à la section 4. À la section 5, nous appliquons notre approche pour produire des estimations de la composition par type de ménages des municipalités basées sur le registre norvégien des ménages, qui présente un sous-enregistrement inférieur au nombre d'identification du logement (NIL). Enfin, à la section 6, nous présentons un résumé.

Estimation de la composition sur petits domaines en présence de données manquantes informatives

Li-Chun Zhang¹

Résumé

L'estimation de la composition sur petits domaines peut poser un problème de données manquantes informatives, si la probabilité que les données manquent varie d'une catégorie d'intérêt à l'autre, ainsi que d'un petit domaine à l'autre. Nous élaborons une approche de modélisation mixte double qui combine un modèle mixte à effets aléatoires pour les données complètes sous-jacentes et un modèle mixte à effets aléatoires du mécanisme de création différentielle de données manquantes. L'effet du plan d'échantillonnage peut être intégré au moyen d'un modèle d'échantillonnage sous quasi-vraisemblance. L'erreur quadratique moyenne conditionnelle de prédiction associée est approximée sous forme d'une incertitude hypothétique de l'estimation des paramètres basée sur les données complètes latentes et une autre correction de composition en trois parties, correspondant à une variance de prédiction *naïve*, une correction positive qui tient compte de la vraisemblance pour la variation supplémentaire due aux données manquantes. Nous illustrons notre approche en l'appliquant à l'estimation de la composition des ménages des municipalités au moyen des données sur les ménages tirées des registres norvégiens, qui présentent un sous-enregistrement informatif du numéro d'identification du logement.

Mots clés : EQM conditionnelle de prédiction ; algorithme EMQVP ; estimation avec préservation des structures généralisée (SPREE généralisée) ; ne manquant pas au hasard ; tableau de contingence à double entrée.

1. Introduction

Des chiffres de population sur petits domaines (ou régions) en fonction de diverses caractéristiques socio-économiques sont demandés de plus en plus fréquemment pour l'affectation de fonds, la planification régionale et la recherche socioéconomique. Purcell et Kish (1980) ont décrit la méthode dite d'estimation avec préservation des structures (SPREE pour Structure Preserving Estimation), dont le fonctionnement consiste à modifier les estimations sur petits domaines de façon qu'elles varient d'un domaine à l'autre conformément à la variation qui existe dans un autre tableau auxiliaire connu de la même variable. Habituellement, le tableau auxiliaire est tiré d'un recensement antérieur ou d'un registre administratif contenant des renseignements similaires. Zhang et Chambers (2004) ont élaboré une approche SPREE généralisée (GSPREE). Ils ont intégré des modèles mixtes à effets fixes ainsi qu'à effets aléatoires et ont montré que le modèle log-linéaire restreint qui sous-tend l'estimation SPREE est un cas particulier. Cette approche offre un moyen de réduire le biais éventuel des estimations SPREE traditionnelles. Le lecteur consultera (1999) pour d'autres approches bayésiennes hiérarchiques et empiriques applicables à ce type de données.

Dans le présent article, nous étendons l'approche GSPREE à des situations dans lesquelles des données manquent. Cette extension peut être utile dans le cas d'enquêtes par sondage pour lesquelles la non-réponse est inévitable. Nous nous intéressons tout spécialement aux

Il convient de souligner que les bureaux nationaux de la statistique qui mènent de grandes enquêtes tiennent compte des données manquantes par répondération ou par imputation. Toutefois, cette correction est faite à des niveaux d'agrégation significativement plus élevés que les petits

Lissage est décrite à la section 2.

crée de données manquantes. L'approche de double et un modèle mixte à effets aléatoires pour le mécanisme de modélisation mixte double qui combine un modèle mixte observés étaient complètes. Nous proposons une approche biais si l'estimation est effectuée comme si les données fausses les données complètes sous-jacentes et produit un différent en ce qui concerne les données manquantes de données manquantes varie selon le domaine. Cette manquant au hasard (Rubin 1976). En outre, le taux global catégorie d'intérêt à l'autre. Il ne s'agit donc pas de données manquantes est *informatif* à condition qu'il varie d'une nous disons que le mécanisme de création des données, dans le contexte de la composition sur petits domaines, total de personnes de 16 à 74 ans appartenant à ce domaine.

compositions sur petits domaines qui peuvent être disposées dans un tableau à double entrée, où l'une des deux dimensions renvoie aux petits domaines et l'autre, aux catégories d'intérêt. La somme des fréquences par cellule est égale à un total de domaine fixe qui peut être connu ou non. Par exemple, chaque personne de 16 à 74 ans peut être classée selon sa situation d'activité, c'est-à-dire « occupée », « en chômage » ou « inactive ». La somme de ces trois chiffres à l'intérieur d'un petit domaine est égale au nombre total de personnes de 16 à 74 ans appartenant à ce domaine.

- Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. et Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-176 (avec discussion).
- Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada : évaluation et application. *Techniques d'enquête*, 27, 69-79.
- Gurney, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics*, American Statistical Association, 242-257.
- Hansen, M.H., Hurwitz, W.N. et Meadow, W.G. (1953). *Sample survey methods and theory*. 2. New York : John Wiley & Sons, Inc.
- Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge : Cambridge University Press.
- Harvey, A.C., et Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.
- Holbrook, A.L., Green, M.C. et Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly*, 67, 79-125.
- Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Section on Social Statistics*, American Statistical Association, 300-303.
- Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.
- Koopman, S.J., Shephard, N. et Doornik, J.A. (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*. Londres : Timberlake Consultants Press.
- Kumar, S., et Lee, H. (1983). Évaluation de l'application d'estimateurs composites à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.
- Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- Nieuwenbroek, N., et Boonstra, H.J. (2002). Bascula 4.0 reference manual, BPA nt : 279-02-TMO. Statistics Netherlands, Heerlen.
- Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.
- Pfeffermann, D., et Rubin-Bleuer, S. (1993). Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions. *Techniques d'enquête*, 19, 159-174.
- Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.
- Finley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. et Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-176 (avec discussion).
- Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada : évaluation et application. *Techniques d'enquête*, 27, 69-79.
- Gurney, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics*, American Statistical Association, 242-257.
- Hansen, M.H., Hurwitz, W.N. et Meadow, W.G. (1953). *Sample survey methods and theory*. 2. New York : John Wiley & Sons, Inc.
- Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge : Cambridge University Press.
- Harvey, A.C., et Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.
- Holbrook, A.L., Green, M.C. et Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly*, 67, 79-125.
- Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Section on Social Statistics*, American Statistical Association, 300-303.
- Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.
- Koopman, S.J., Shephard, N. et Doornik, J.A. (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*. Londres : Timberlake Consultants Press.
- Kumar, S., et Lee, H. (1983). Évaluation de l'application d'estimateurs composites à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.
- Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- Nieuwenbroek, N., et Boonstra, H.J. (2002). Bascula 4.0 reference manual, BPA nt : 279-02-TMO. Statistics Netherlands, Heerlen.
- Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.
- Pfeffermann, D., et Rubin-Bleuer, S. (1993). Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions. *Techniques d'enquête*, 19, 159-174.
- Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.
- van der Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch labour force survey. *Journal of the Royal Statistical Society, Series A*, 171, 581-613.
- van den Brakel, J.A., et Krieger, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. Research paper, Statistics Netherlands, Heerlen (<http://www.cbs.nl/en-Gb/menu/methoden/research/discussionpapers/archief/2009/default.htm?Language=switch=on>).
- Statistik Canada, Canada, N° 12-001-X au catalogue

dans les modèles. Si une série plus longue devient disponible, une composante cyclique supplémentaire pourrait être nécessaire pour refléter les fluctuations économiques. Une autre amélioration possible consisterait à détecter et à modéliser les valeurs aberrantes. En outre, le modèle doit être étendu à l'estimation des taux de chômage mensuel pour différents domaines en utilisant l'information d'échantillon recueillie par le passé, ainsi que les données transversales provenant d'autres petits domaines, en suivant l'approche proposée par Pfeffermann et Burck (1990) et par Pfeffermann et Tillier (2006).

Remerciements

Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas forcément la politique de Statistics Netherlands. Les auteurs remercient le professeur D. Pfeffermann et le professeur S.J. Koopman de leurs conseils précieux durant le projet, ainsi que le rédacteur associé et les examinateurs de leurs commentaires constructifs au sujet d'ébauches antérieures du présent article.

Bibliographie

Baillar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Bell, W.R., et Hillmer, S.C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Binder, D.A., et Dick, J.P. (1989). Enquêtes répétées - Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.

Binder, D.A., et Dick, J.P. (1990). Méthode pour l'analyse des modèles ARIMA. *Techniques d'enquête*, 16, 251-265.

Box, G.E.P., et Jenkins, G.W.M. (1970). *Time series analysis - forecasting and control*. San Francisco : Holden-Day.

Cantwell, P.J. (1990). Formules de variance pour l'estimateurs composites dans les plans de renouvellement. *Techniques d'enquête*, 16, 163-174.

Doomik, J.A. (1998). *Object-oriented matrix programming using Ox 2.0*. Londres : Timberlake Consultants Press.

Durbin, J., et Koopman, S.J. (2001). *Time series analysis by state space methods*. Oxford : Oxford University Press.

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55, 182-199.

sous le plan avec renouvellement de panel de l'EPA illustre clairement l'existence d'erreurs non dues à l'échantillon-nage, telles que des erreurs de mesure et celles dues à l'érosion du panel. Par conséquent, les concepts classiques, selon lesquels les observations obtenues auprès des unités échantillonnées sont les valeurs fixes réelles observées sans erreur et les répondants peuvent être considérés comme un échantillon probabiliste représentatif de la population cible, qui sont généralement appliqués en théorie de l'échantillonage fondé sur le plan de sondage ne tiennent pas sous ce genre de plan rotatif. L'application d'estimateurs directs en cas d'erreurs de mesure et d'érosion sélective du panel donnera lieu à des estimations gravement biaisées. Dans la procédure d'estimation ordinaire, on applique aux estimations par régression généralisée une correction par le ratio qui est basé sur l'hypothèse de modélisation implicite que le biais est constant sur une période de trois ans. Le modèle de séries chronologiques appliqué dans la présente étude peut être utilisé pour produire des estimations corrigées de manière plus poussée du biais introduit par ces erreurs non dues à l'échantillonage.

Cette méthode d'estimation est également applicable dans des situations où la petite taille des échantillons donne des erreurs-types inacceptablement grandes. Des petites tailles d'échantillon se produisent si des statistiques officielles sont requises pour de petits domaines ou pour de courtes périodes de collecte des données, comme cela est le cas des taux de chômage mensuels dans l'EPA. La plupart des enquêtes réalisées par les instituts nationaux de statistique sont effectuées en temps continu et sont fondées sur des plans de sondage transversaux ou avec renouvellement de panel. Par conséquent, les méthodes d'estimations basées sur des modèles de séries chronologiques qui utilisent l'information recueillie auprès d'échantillons observés aux périodes précédentes sont particulièrement intéressantes.

Le modèle de séries chronologiques produit des estimations de la tendance et des composantes saisonnières du paramètre de population. Les estimations désaisonnalisées du paramètre et leurs erreurs d'estimation sont par conséquent obtenues en tant que sous-produit de cette méthode d'estimation. Un autre avantage important est que cette approche tient compte de l'autocorrélation entre les erreurs d'enquête dues au plan avec renouvellement de panel. Pfeffermann et coll. (1998) montrent que le fait d'ignorer ces autocorrélations, par exemple en se servant de filtres de Henderson dans X-11-ARIMA (Findley, Monsell, Bell, Otto et Chen 1998), produit des estimations fausses de la

tendance.

Le modèle peut être amélioré de plusieurs façons. L'information sur le chômage enregistré et les variables connexes, disponibles dans le registre du bureau de l'emploi et du revenu, peut être utilisée comme variable auxiliaire

L'hypothèse de modélisation selon laquelle les estimations basées sur les données de la première vague sont sans biais, le modèle de séries chronologiques qui tient compte du BGR dans les effets saisonniers est celui que nous privilégions, puisque qu'il élimine le biais dans la composante saisonnière.

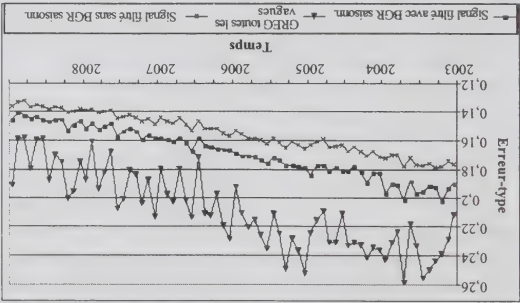


Figure 5.12 Erreurs-types des estimations par régression généralisée (GREG) basées sur toutes les vagues et des estimations filtrées pour deux modèles distincts de séries chronologiques pour le taux de chômage mensuel

Discussion

Le présent article décrit l'application aux données mensuelles de l'EPA d'un modèle structural multivarié de séries chronologiques qui tient compte du plan avec renouvellement de panel de l'enquête. Cette approche, proposée pour la première fois par Pfeffermann (1991), est étendue ici au moyen d'une composante qui modélise les différences systématiques d'effets saisonniers entre les vagues successives. Comparativement à l'estimateur par régression généralisée, qui est utilisé à l'heure actuelle pour produire les estimations ordinaires de l'EPA, le modèle de séries chronologiques donne des estimations du taux de chômage d'une précision considérablement plus grande. En premier lieu, ce modèle estime explicitement le biais de groupe de renouvellement (BGR) dans la tendance et la composante saisonnière observée entre la première vague de collecte par PPAO et les quatre vagues successives de collecte par ITAO. Deuxièmement, il est renforcé par des données observées aux périodes précédentes par la voie du modèle hypothétique utilisé pour le paramètre de population et de l'autocorrélation entre les erreurs d'enquête des divers panels.

Le BGR induit par le plan avec renouvellement de panel est considérable. Le biais dans la tendance donne lieu à une sous-estimation du taux de chômage pour les vagues subséquentes à la première et sa grandeur diminue légèrement, pour passer de -0,8 point de pourcentage à la deuxième vague à -1,1 point de pourcentage à la cinquième.

Les composantes saisonnières des deux premières vagues et des trois dernières diffèrent également de manière significative, parce que les effets saisonniers sont moins prononcés dans les trois dernières vagues.

L'utilisation d'un modèle de séries chronologiques parcomposante saisonnière, qui tient compte du BGR dans la tendance, mais non du BGR dans la composante saisonnière donne lieu à une réduction supplémentaire de l'erreur-type des estimations filtrées. Cependant, cette simplification du modèle introduit un biais dans la composante saisonnière dans les estimations mensuelles du taux de chômage. Puisque les erreurs-types des estimations filtrées obtenues sous ce modèle parcomposante ne reflètent pas ce biais, nous préférons utiliser un modèle de séries chronologiques qui tient compte du BGR à la fois dans la tendance et dans la composante saisonnière.

Le modèle de séries chronologiques est déterminé en imposant une contrainte au BGR des paramètres, c'est-à-dire en supposant que la première vague est observée sans biais. Autrement dit, les estimations fondées sur les données de la première vague sont utilisées pour étalonner les estimations des vagues subséquentes. Si l'on applique cette contrainte, il est nécessaire d'essayer par tous les moyens, à chaque étape du processus statistique, de réduire le biais éventuel dans les données de la première vague, par exemple en utilisant le mode de collecte de données le plus approprié, en réduisant la non-réponse et en optimisant le schéma de pondération. Selon l'information externe au sujet du biais aux diverses vagues, il est possible d'ajuster les contraintes concernant les composantes du BGR.

L'approche de séries chronologiques examinée ici convient pour produire des estimations fondées sur un modèle du taux de chômage mensuel. Cependant, Statistics Netherlands se montre généralement assez réservé en ce qui concerne l'application de méthodes d'estimations fondées sur un modèle pour la production de statistiques officielles. La spécification incorrecte du modèle pourrait aboutir à des estimations présentant un biais important. Ce biais n'est pas reflété dans les erreurs-types des estimations par filtre de Kalman. Par conséquent, de nombreux efforts de sélection et d'évaluation du modèle sont requis pour chaque variable cible, ce qui rend difficile une application directe de ce genre de techniques d'estimation, puisque le temps disponible pour la phase d'analyse du processus de production ordinaire de statistiques officielles est généralement limité.

Par ailleurs, il existe des arguments en faveur de la production de séries de données officielles s'appuyant sur des procédures fondées sur un modèle assorties d'une méthodologie et de descriptions de qualité appropriées dans des situations où les estimateurs directs ne fournissent pas d'estimations suffisamment fiables. Le BGR observé

Comme prévu, les erreurs-types des estimations par régression généralisée basées sur toutes les vagues sont plus faibles que celles des estimations par régression généralisée basées sur la vague d'IPAO puisque ces estimations s'appuient sur un plus grand nombre de données. Les erreurs-types des estimations par régression généralisée basées sur toutes les vagues, parce que le modèle de séries chronologiques utilise de l'information supplémentaire provenant des périodes précédentes. Les erreurs-types des estimations filtrées sont légèrement, mais continuellement, décroissantes au cours de la période allant de 2003 à 2008.

La taille et la complexité du modèle de séries chronologiques appliqué sont grandes comparativement à la longueur de la série disponible pour ajuster le modèle. Le modèle final qui est appliqué à une série pentadimensionnelle observée mensuellement durant une période de huit ans contient 41 variables d'état. Par conséquent, cela vaut la peine de considérer des modèles plus parcimonieux, susceptibles de réduire les erreurs-types des estimations filtrées. De surcroît, l'estimation par régression généralisée présente un biais, car le BGR contient un effet saisonnier qui n'est pas

Figure 5.10 Erreurs-types des estimations par régression généralisée (GREG) et des estimations filtrées du taux de chômage mensuel

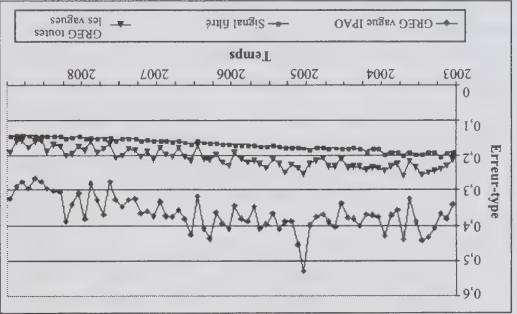


Figure 5.9 Estimations par régression généralisée (GREG) du taux de chômage mensuel basées sur la vague IPAO et basées sur toutes les vagues

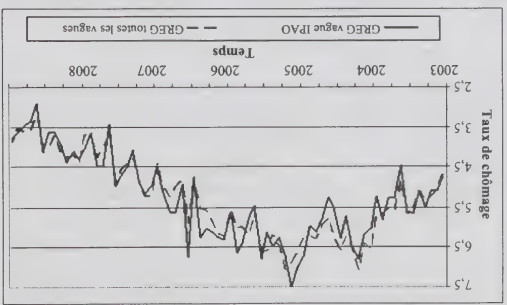
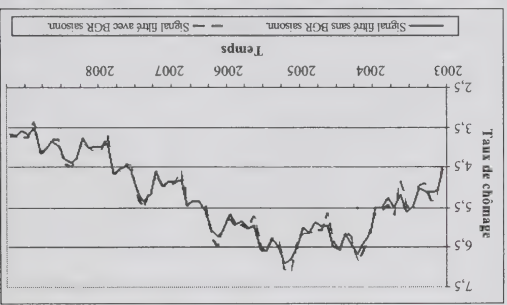


Figure 5.11 Estimations filtrées du taux de chômage mensuel pour deux modèles de séries chronologiques différents



composante saisonnière sont comparées à la figure 5.11. Le modèle sans composante pour le BGR des effets saisonniers repose sur l'hypothèse d'un effet saisonnier pour le paramètre de population θ , qui est basé sur une moyenne des effets saisonniers des cinq vagues. Les valeurs absolues faibles sous le modèle simplifié, ce qui donne une estimation plus faible du taux de chômage mensuel en février et en mars, et une estimation plus grande en août. Cela se traduit par un profil saisonnier plus prononcé dans la série filtrée obtenue au moyen du modèle complet.

Les erreurs-types des estimations filtrées obtenues avec les deux modèles de séries chronologiques et les erreurs-types des estimations par régression généralisée en utilisant les données de toutes les vagues sont comparées à la figure 5.12. L'erreur-type des estimations filtrées du modèle de séries chronologiques simplifié est considérablement plus faible que celle des estimations par régression généralisée calculées en se servant des données de toutes les vagues. La simplification du modèle de séries chronologiques en omettant le BGR pour les effets saisonniers produit une réduction supplémentaire de l'erreur-type au prix d'un accroissement du biais dans les effets saisonniers. Sous

À la figure 5.7, nous comparons les estimations par régression généralisée du taux de chômage mensuel basées sur les données de la vague d'IPAO aux estimations filtres saisonniers sous le modèle de séries chronologiques. Il s'ensuit également que les estimations filtres sont corrigées du BGR, puisque la série filtree est au même niveau que la série d'estimations par régression généralisée basées sur les données de la vague IPAO. Il en est ainsi parce que nous émettons l'hypothèse que les paramètres du modèle pour le BGR pour la première vague sont nuls (section 3.2). Cela implique que les vagues d'ITAO sont étalonnées d'après les résultats de la première vague.

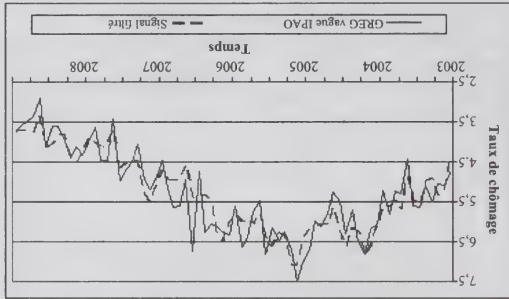


Figure 5.7 Estimations filtres et estimations par régression généralisée (GREG) basées sur la vague d'IPAO du taux de chômage mensuel

La méthode appliquée dans la procédure ordinaire d'estimation de l'IPA pour combiner les vagues d'ITAO et d'IPAO est également utilisée pour estimer les taux de chômage mensuels. À la figure 5.8, nous comparons les estimations par régression généralisée du taux de chômage mensuel basées sur les cinq vagues en utilisant la formule (2.4) aux estimations filtres. Les deux estimations du taux de chômage mensuel ont le même niveau, puisqu'elles sont toutes les deux étalonnées d'après les résultats de la première vague. L'estimateur par la régression généralisée est étalonné de manière assez rigide en utilisant le ratio (2.3), qui est supposé d'avance comme étant constant sur une période de trois ans. Les estimations filtres sont étalonnées d'une manière plus subtile par modélisation explicite de la tendance et de la saisonnalité dans le BGR. La présence de cette saisonnalité indique que l'hypothèse d'un BGR constant ne tient pas. À la figure 5.9, nous comparons également les estimations par régression

généralisée mensuelles basées sur toutes les vagues aux estimations par régression généralisée basées sur la vague

d'IPAO.

La correction par le ratio appliquée dans la formule (2.4) aux estimations par régression généralisée basées sur toutes les vagues élimine le BGR dans la tendance, mais ne corrige pas le BGR dans les effets saisonniers, comme le montre les figures 5.8 et 5.9. La série d'estimations par régression généralisée basées sur toutes les vagues suit la même courbe que les estimations par régression généralisée basées sur la vague d'IPAO (figure 5.9). Cependant, il existe des différences subtiles entre les estimations filtres et les estimations par régression généralisée basées sur toutes les vagues (figure 5.8). Ces différences sont dues en partie au fait que certains creux et pics dans la série d'estimations par régression généralisée sont considérés comme des erreurs d'enquête par le modèle de séries chronologiques, et qu'elles sont aussi le résultat de différences systématiques entre les effets saisonniers des vagues successives. Par exemple, les estimations du modèle sont plus grandes en février et en mars 2003, 2005 et 2006, et plus petites en août en 2004, 2005 et 2006.

À la figure 5.10, nous comparons les erreurs-types des estimations par régression généralisée basées sur toutes les vagues, des estimations par régression généralisée basées sur la vague d'IPAO et des estimations filtres. Les erreurs-types des estimations par régression généralisée sont calculées comme il est décrit à la section 2.4. Celles des estimations filtres sont obtenues au moyen des formules de récursion classique du filtre de Kalman; voir Harvey (1989) ou Durbin et Koopman (2001). Les récursions du filtre de Kalman reposent sur l'hypothèse que le modèle d'espace d'états ajusté correspond à la situation réelle. Par conséquent, les erreurs-types des estimations filtres ne reflètent pas la variation supplémentaire induite par l'utilisation des estimations de la vraisemblance pour les composantes de la variance dans le modèle d'espace d'états et sont, par conséquent, trop optimistes.

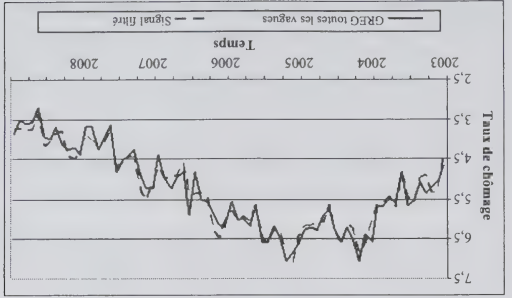


Figure 5.8 Estimations filtres et estimations par régression généralisée (GREG) basées sur toutes les vagues du taux de chômage mensuel

Un résultat empirique intéressant de la présente application est la constatation que le BGR présente une saisonnalité. Les estimations par filtre de Kalman du BGR des effets saisonniers sont également temporellement in-variantes. Par conséquent, nous effectuons une série de tests du rapport de vraisemblance pour atteindre le modèle choisi en dernière analyse et pour déterminer si les effets saisonniers dans le BGR de ce modèle sont collectivement significativement différents de zéro. Considérons les modèles emboîtés suivants :

- M1 : BGR distinct et fixe dans la saisonnalité pour les deuxième, troisième, quatrième et cinquième vagues ;
- M2 : même que M1, avec le BGR dans la saisonnalité de la deuxième vague égal à zéro ;
- M3 : même que M2 avec BGR égal dans la saisonnalité des troisième, quatrième et cinquième vagues ;
- M4 : BGR dans la saisonnalité des deuxième, troisième, quatrième et cinquième vagues égal à zéro.

Les résultats des tests du rapport de vraisemblance de cette série de modèles sont décrits au tableau 5.3.

Tableau 5.3
Tests du rapport de vraisemblance pour le BGR dans la saisonnalité

Modèle	Log-vraisemblance	Hypothèse nulle	Statistique du rapport de vraisemblance	ddl	Valeur p
M1	1 592,9			11	0,19568
M2	1 585,5		M2 = M1	14,7	0,03422
M3	1 573,7		M3 = M2	23,7	0,00373
M4	1 559,9		M4 = M3	27,6	

Le test de vérification de l'hypothèse que M2 est égal à M1 montre que la saisonnalité de la deuxième vague ne diffère pas de manière significative de celle de la première vague. Le test de vérification de l'hypothèse que M3 est égal à M2 montre que les BGR dans la saisonnalité des troisième, quatrième et cinquième vagues ne diffèrent pas de manière significative. La vérification de l'hypothèse que M4 est égal à M3 montre que les BGR des effets saisonniers dans les trois dernières vagues sont conjointement significativement différents de zéro.

Les estimations lissées par filtre de Kalman du BGR des effets saisonniers pour les troisième, quatrième et cinquième vagues sont présentées à la figure 5.5. Les estimations lissées par filtre de Kalman des effets saisonniers sont comparées aux estimations lissées du BGR des effets saisonniers à la figure 5.6.

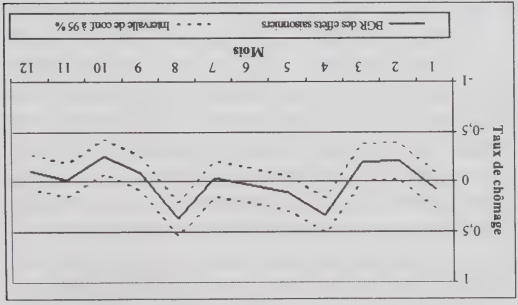


Figure 5.5 Estimations lissées par filtre de Kalman du BGR des effets saisonniers dans les troisième, quatrième et cinquième vagues

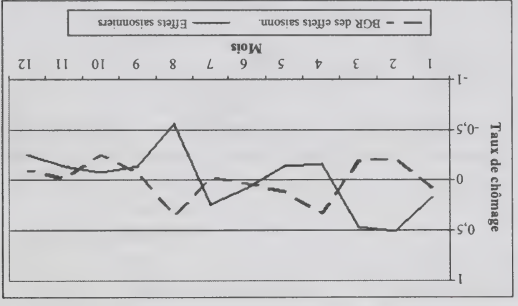


Figure 5.6 Comparaison des estimations lissées par filtre de Kalman du BGR des effets saisonniers dans les troisième, quatrième et cinquième vagues et des effets saisonniers en 2008

Il découle de la figure 5.5 que les effets saisonniers en février, mars, avril, août et octobre durant les troisième, quatrième et cinquième vagues divergent de manière significative de ceux des première et deuxième vagues. La figure 5.6 montre que le BGR des effets saisonniers annule en grande en grande partie ces effets durant ces mois. Les effets saisonniers aux trois dernières vagues semblent moins prononcés que dans les deux premières. Les divers facteurs qui contribuent au BGR dans la tendance ainsi que dans les effets saisonniers sont résumés à la section 2.2.

5.3 Comparaison avec les estimations par régression généralisée

À la présente section, nous comparons les estimations par régression généralisée mensuelles du taux de chômage et les estimations lissées par filtre de Kalman du BGR des effets saisonniers pour le mois 7.

5.2 Résultats d'estimation pour le modèle de séries

chronologiques

Les estimations du maximum de vraisemblance des hyperparamètres, c'est-à-dire les composantes de la variance des processus stochastiques pour les variables d'état, sont obtenues en utilisant une méthode d'optimisation numérique (algorithme BFGS, Doornik 1998). Pour éviter d'obtenir des estimations de variance négatives, nous avons estimé les variances log-transformées. Les estimations du maximum de vraisemblance de la variance log-transformée du niveau de la tendance (σ^2_η), de la composante saisonnière (σ^2_ω), du BGR de la tendance (σ^2_γ) et du BGR des composantes saisonnières (σ^2_δ) tend vers de grandes valeurs négatives avec des erreurs-types extrêmement grandes. La valeur de ces composantes de la variance est par conséquent mise à zéro dans le modèle final. Les résultats d'estimation des autres hyperparamètres sont présentés au tableau 5.1.

Tableau 5.1

Estimation du maximum de vraisemblance des hyperparamètres

Hyperparamètre	Comp. de la variance	Estimation	Erreur-type	Intervalle de confiance à 95 %	Borne inf.	Borne sup.	Pente (σ^2_η)
							Comp. irrégulière (σ^2_ϵ)
							-17,226
							0,549
							0,482
							-13,480
							1,183E-3
							0,737E-3
							1,897E-3

Les estimations lissées par filtre de Kalman du taux de chômage θ_t sont données à la figure 5.2. Il s'agit des estimations du taux de chômage mensuel basées sur le modèle de la tendance lisse et une composante saisonnière, corrigées du BGR entre les cinq estimations par régression généralisée. Le modèle de tendance lisse locale est simplifié en un modèle de tendance lisse puisque $\sigma^2_\gamma = 0$. La composante tendance varie en fonction du temps puisque l'estimation du maximum de vraisemblance de l'hyperparamètre de la pente est positive (voir le tableau 5.1). La composante saisonnière dépend également du temps, puisque $\sigma^2_\omega = 0$. Par conséquent, les estimations des effets saisonniers obtenues au moyen de la forme trigonométrique sont exactement les mêmes que les résultats obtenus avec le modèle saisonnier à variables indicatrices bien connu. Les estimations de la tendance et de la composante saisonnière lissées par le filtre de Kalman sont représentées graphiquement aux figures 5.3 et 5.4, respectivement.

Les estimations par filtre de Kalman du BGR de la tendance sont indépendantes du temps. Les estimations lissées par filtre de Kalman du BGR sont données au tableau 5.2. Le modèle détecte merveilleusement bien un biais légèrement croissant dans la tendance des vagues

successives. Les estimations du BGR pour les quatre vagues d'ITAO sont significativement différentes de zéro.

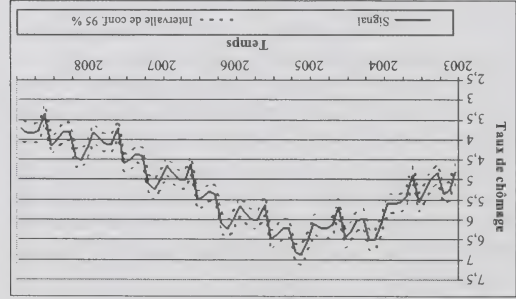


Figure 5.2 Estimations lissées par filtre de Kalman du taux de chômage mensuel

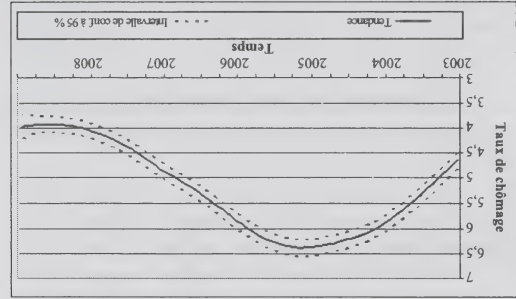


Figure 5.3 Estimations lissées par filtre de Kalman de la tendance du taux de chômage mensuel

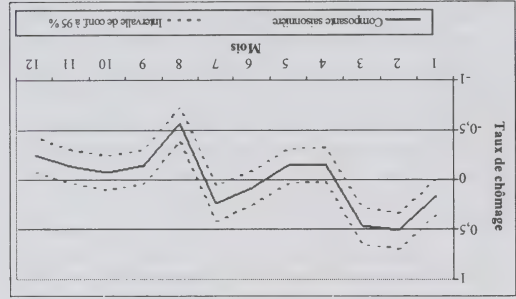


Figure 5.4 Estimations lissées par filtre de Kalman de la composante saisonnière du taux de chômage mensuel

Tableau 5.2

Estimations lissées par filtre de Kalman du BGR de la tendance

Vague	BGR	Erreur-type
2	-0,75	0,04
3	-0,86	0,04
4	-0,96	0,05
5	-1,10	0,05

d'améliorer la précision de l'estimation du taux de chômage mensuel. Les composantes pour le BGR, élaborées à la section 3.2, tiennent compte des écarts systématiques entre les cinq estimations par régression généralisée mensuelles, afin d'éviter que l'estimation du taux de chômage mensuel soit entachée de ce biais. La composante pour les erreurs d'enquête, élaborée à la section 3.3, tient compte de l'auto-corrélation entre les cinq estimations par régression généralisée basées sur le même échantillon, observées à intervalles trimestriels. Bien que cette approche soit fondée sur un modèle, elle tient compte de la complexité du plan de sondage de l'EPA, puisque les estimations par régression généralisée sont utilisées comme données d'entrée.

4. Représentation en espace d'états

Le modèle de séries chronologiques pour les cinq estimations par régression généralisée mensuelles élaborées à la section 3 peut être exprimé dans la représentation en espace d'états ; voir Harvey (1989) ou Durbin et Koopman (2001). Un modèle à espace d'états est constitué d'une équation de mesure et d'une équation de transition. L'équation de mesure, qui est parfois également appelée équation de signal, spécifie comment les observations dépendent d'une combinaison linéaire du vecteur d'états qui contient les variables d'état inobservées pour la tendance, la composante saisonnière, le BGR et les erreurs d'enquête. L'équation de transition, qui est parfois également appelée équation du système, spécifie comment le vecteur d'états évolue en fonction du temps. La représentation en espace d'états du

et Krieg (2009).

Sous l'hypothèse que les termes d'erreur suivent une loi normale, nous pouvons appliquer le filtre de Kalman pour obtenir les estimations optimales du vecteur d'états. Les estimations pour les variables d'état pour la période t basées sur l'information disponible jusqu'à la période t inclusivement sont appelées estimations filtrées. Les estimations filtrées des vecteurs d'états antérieurs peuvent être mises à jour si de nouvelles données deviennent disponibles. Cette procédure porte le nom de lissage et donne des estimations lissées qui sont fondées sur la séries chronologiques entièrement observée. Donc, l'estimation lissée pour le vecteur d'états pour la période t tient compte également de l'information devenue disponible après la période t . Dans le présent article, les estimations du filtre de Kalman pour les variables d'état sont lissées au moyen du lisseur à intervalle fixe. Voir Harvey (1989), ainsi que Durbin et Koopman (2001) pour des précisions techniques.

L'analyse est effectuée au moyen d'un logiciel développé en langage Ox en combinaison avec les sous-routines de Ssfpack 3.0 ; voir Doornik (1998), ainsi que Koopman,

5.1 Analyses préliminaires

5. Résultats

Shephard et Doornik (2008). Toutes les variables d'état sont non stationnaires à l'exception des erreurs d'enquête. Les variables non stationnaires sont initialisées au moyen d'un prior diffus, autrement dit les espérances des états initiaux sont nulles et la matrice de covariance initiale des états est diagonale avec de grands éléments diagonaux. Les erreurs d'enquête sont stationnaires et, par conséquent, sont initialisées en se servant d'un prior approprié. Les valeurs initiales des erreurs d'enquête sont nulles et la matrice de covariance des erreurs d'enquête est la section 3.3. Dans Ssfpack 3.0, une fonction de log-vraisemblance diffuse exacte est obtenue à l'aide de la procédure proposée par Koopman (1997).

En nous servant de l'estimateur par régression généralisée, nous avons obtenu des estimations mensuelles du taux de chômage pour chaque vague de la manière décrite à la section 2.4. À la figure 5.1, nous comparons le taux de chômage fondé sur la vague d'ITAO à la moyenne des taux obtenus pour les quatre vagues d'ITAO. Le graphique révèle que le taux de chômage observé pour la première vague est systématiquement plus élevé que celui obtenu pour les quatre autres.

Nous avons modélisé les cinq séries chronologiques obtenues pour les diverses vagues en nous servant du modèle de séries chronologiques proposés aux sections 3 et 4. Les analyses préliminaires indiquent que les estimations pour le BGR des effets saisonniers à la deuxième vague ne diffèrent pas de manière significative de zéro et que les BGR pour les effets saisonniers aux troisième, quatrième et cinquième vagues ne diffèrent pas de manière significative l'un de l'autre. Par conséquent, nous avons simplifié le modèle pour qu'il ne contienne qu'un seul effet saisonnier avec BGR. Voir Van den Brakel et Krieg (2009) pour la représentation en espace d'états.

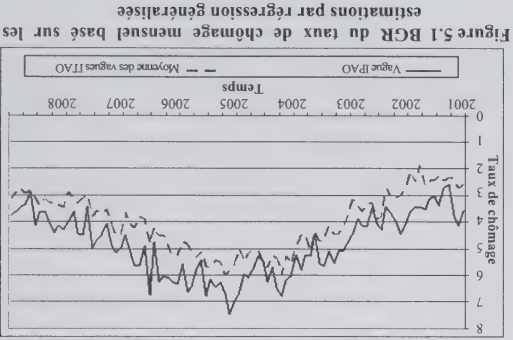


Figure 5.1 BGR du taux de chômage mensuel basé sur les estimations par régression généralisée

Tableau 3.1
Corrélations et autocorrélations partielles des erreurs d'enquête
des divers panels

Vague		D��calage			
		1	2	3	4
1	AC	-0,029	0,264	0,022	0,230
	ACP	-0,029	0,263	0,038	0,175
2	AC	0,291	0,135	0,035	-0,250
	ACP	0,291	0,054	-0,020	-0,287
3	AC	0,240	0,120	0,087	0,219
	ACP	0,240	0,066	0,047	0,194
4	AC	0,442	0,253	0,122	0,156
	ACP	0,442	0,072	-0,016	0,115
5	AC	0,249	0,298	-0,183	0,127
	ACP	0,249	0,252	-0,344	0,218
Moyenne*	AC	0,306	0,224	-0,030	0,127
	ACP	0,306	0,144	-0,150	0,162

Les valeurs soulignées des AC et des ACP correspondent à des vagues avec chevauchement d'échantillons
*: Les moyennes sont fondées sur les vagues avec chevauchement d'échantillons

Les erreurs-types des AC estimées sont égales à $1/\sqrt{T}$, où T désigne le nombre d'observations. Cela implique que les corrélations dont la valeur absolue est supérieure à 0,21 diffèrent de manière significative de zéro au seuil de 5 %. Dans le tableau 3.1, les décalages ont trait à des périodes de trois mois, de sorte que le décalage unitaire est égal à un décalage temporel de trois mois, le décalage deux, à un décalage temporel de six mois, etc.

Les échantillons chevauchants sont significativement différents de zéro pour le décalage 1 pour les échantillons chevauchants et pour toutes significativement différentes de zéro. Pour le décalage 2, les AC pour les échantillons chevauchants sont significativement différentes de zéro pour les quatrièmes et cinquièmes vagues, mais non pour la troisième. Les AC qui sont fondées sur des échantillons non chevauchants sont parfois d'une grandeur inattendue, par exemple telles observées pour les décalages 2 et 4 de la première vague, et le décalage 4 de la troisième vague. Par ailleurs, l'AC pour le décalage 4 de la deuxième vague possède un

Pfeffermann et coll. (1998) signalent aussi une grande valeur positive de AC pour les décalages avec des échantillons non chevauchants. Dans leur cas, cela peut s'expliquer par le fait que les échantillons sont remplacés dans les petites régions géographiques. Dans l'échantillon de l'EPA des Pays-Bas, le remplacement a lieu au niveau national. Il n'existe aucune bonne raison pour laquelle les AC pour les échantillons non chevauchants sont parfois petites et prennent parfois des valeurs importantes positives ainsi que

Troisième, quatrième et cinquième vagues.

Les estimations directes de la variance et de la structure de covariance des erreurs d'enquête sont combinées dans le modèle de séries chronologiques en utilisant la forme générale $e'_{t-j} = k'_{t-j}e'_t$ du modèle d'erreur d'enquête, où $k'_{t-j} = \sqrt{\text{Var}(x'_{t-j})}$; voir Binder et Dick (1990). Cela permet de tenir compte de la non-homogénéité de la variance des erreurs d'enquête causée, par exemple, par la diminution progressive de la taille de l'échantillon au cours de la dernière décennie.

Puisque la première vague n'est pas correlée aux erreurs d'enquête obtenues par le passé, nous supposons que e'_t est un bruit blanc avec $E(e'_t) = 0$ et $\text{Var}(e'_t) = 1$. Par conséquent, la variance de l'erreur d'enquête est égale à $\text{Var}(e'_t) = (k'_t)^2$, qui est égale à l'estimation directe de la variance de l'estimation par régression généralisée pour la première vague. Pour les deuxième, troisième, quatrième et cinquième vagues, nous supposons que $e'_{t-j} = p e'_{t-j-1} + v'_{t-j}$, avec $p = 0,306$, et

$$E(v'_{t-j}) = 0, \text{Cov}(v'_{t-j}, v'_{t-j-1}) = 0^v, \begin{cases} 0 & \text{si } t \neq t' \\ 0^v & \text{si } t = t' \end{cases}$$

Puisque e'_{t-j} est un processus AR(1), $\text{Var}(e'_{t-j}) = \sigma_v^2 / (1 - p^2)$. Pour que $\text{Var}(e'_{t-j})$ soit égale à l'estimation directe de la variance de l'estimation par régression généralisée, il s'ensuit que $\sigma_v^2 = (1 - p^2)$.

3.4 Modèle de séries chronologiques final pour le taux de chômage mensuel

Nous obtenons le modèle de séries chronologiques pour le vecteur avec l'estimation par régression généralisée Y_t en instaurant dans (3.1) les diverses composantes établies aux sections 3.1 à 3.3. Ce modèle utilise comme données d'entrée les cinq estimations par régression généralisée mensuelles pour obtenir des estimations fondées sur un modèle du taux de chômage mensuel. La composante correspondant au paramètre de population θ_t dans (3.2) que nous avons élaborée à la section 3.1, tire parti de l'information d'échantillon observée dans le passé en vue

estimations de la variance pour les panels individuels au moyen d'un modèle de régression linéaire $\text{Var}(Y_{t-j}^{t-j}) = b_0^j + b_1^j(Y_{t-j}^{t-j}/n_{t-j}^{t-j}) + \text{erreur}$, où n_{t-j}^{t-j} désigne la taille au temps t de l'échantillon qui est entré dans le panel au temps $t - j$.

Le plan de sondage avec renouvellement de panel implique un chevauchement de l'échantillon avec les panels observés dans le passé. L'échantillon de la première vague entre dans le panel pour la première fois au temps t , si bien qu'il n'y a pas de chevauchement avec les panels observés antérieurement. Par conséquent, les erreurs d'enquête de la première vague, e_{t-j}^{t-j} , ne sont pas corrélées avec les erreurs d'enquête antérieures. L'erreur d'enquête de la deuxième vague, e_{t-3}^{t-3} , est corrélée avec l'erreur d'enquête de la première vague, e_{t-6}^{t-6} , est corrélée avec l'erreur d'enquête de la quatrième vague, e_{t-9}^{t-9} , est corrélée avec l'erreur d'enquête de la quatrième vague, e_{t-9}^{t-9} , est corrélée avec l'erreur d'enquête de la cinquième vague, e_{t-12}^{t-12} , est corrélée avec e_{t-12}^{t-12} , e_{t-6}^{t-6} , e_{t-9}^{t-9} et e_{t-12}^{t-12} .

Nous estimons les autocorrélations entre les erreurs d'enquête des vagues successives par la méthode proposée par Pfeffermann et coll. (1998). Puisqu'il est impossible d'observer directement les erreurs d'enquête réelles, cette

approche débute par le calcul des autocovariances pour les pseudo-erreurs d'enquête, qui sont définies comme étant $(Y_{t-j}^{t-j} - \bar{Y}_j)$, où \bar{Y}_j désigne la moyenne des cinq estimations de panel Y_{t-j}^{t-j} au temps t . Les autocovariances des pseudo-erreurs d'enquête pour une vague particulière sont influencées par les autocovariances des erreurs d'enquête réelles

obtenues au temps t . L'équation (4) de Pfeffermann et coll. (1998) spécifie la relation entre les autocovariances des pseudo-erreurs d'enquête et des erreurs d'enquête réelles. De cette équation, il découle que les autocovariances des erreurs d'enquête réelles peuvent être calculées à partir des autocovariances des pseudo-erreurs d'enquête par $\Phi_k = \mathbf{F}^{-1}\mathbf{C}_k$, où \mathbf{C}_k est un vecteur contenant les cinq auto-covariances des pseudo-erreurs d'enquête au décalage k , Φ_k est un vecteur contenant les cinq autocovariances des erreurs d'enquête au décalage k , et \mathbf{F} est une matrice de dimensions $M \times M$ dont les éléments diagonaux sont égaux à $(M - 1/M)^2$. Ici, M désigne le nombre de vagues du plan de sondage par panel ($M = 5$ dans la présente application). Les autocorrélations (AC) et les autocorrélations partielles (ACP) des erreurs d'enquête des vagues successives sont

données au tableau 3.1.

Des successives sont modélisées à l'aide de λ_{t-j}^{t-j} et γ_{t-j}^{t-j} . Des vecteurs sont nécessaires pour spécifier le modèle (3.1). Ici, nous supposons qu'une estimation sans biais de θ_{t-j}^{t-j} est obtenue à la première vague, qui est observée par IPAO, c'est-à-dire Y_{t-j}^{t-j} . Cela implique que la première composante de λ_{t-j}^{t-j} est nulle. Ainsi, λ_{t-j}^{t-j} mesure les écarts systématiques entre les tendances des deuxième, troisième, quatrième et cinquième vagues, et celle de la première vague. Les composantes de λ_{t-j}^{t-j} sont définies comme étant :

$$\lambda_{t-j}^{t-j} = 0, \lambda_{t-j}^{t-j} = \lambda_{t-j-1}^{t-j-1} + n_{t-j-1}^{t-j-1}, j = 3, 6, 9, 12, (3.7)$$

En outre, γ_{t-j}^{t-j} mesure les écarts systématiques des composantes saisonnières par rapport à celle de la première vague. Cela implique que $\gamma_{t-j}^{t-j} = 0$. Les autres composantes de γ_{t-j}^{t-j} sont définies comme des fonctions trigonométriques ayant la forme de (3.5). Nous supposons que la variance des perturbations des composantes saisonnières est la même pour toutes les vagues, et nous la désignons par σ_{γ}^2 .

Afin d'empirer de l'information à toutes les vagues du panel, nous modélisons le BGR de la tendance ainsi que le BGR des composantes saisonnières sous forme de composantes temporellement invariantes, c'est-à-dire $\sigma_{\gamma}^2 = \sigma_{\gamma}^2 = 0$. En tant que sorte de diagnostic du modèle, ce dernier permet initialement que les composantes varient en fonction du temps. Dans cette application, les estimations du maximum de vraisemblance de σ_{γ}^2 et de σ_{γ}^2 sont proches de zéro. S'il n'en est pas ainsi, il pourrait être possible de permettre des composantes du BGR indépendantes du temps pour divers intervalles de temps.

3.3 Modèle de séries chronologiques pour les erreurs d'enquête

Enfin, nous élaborons un modèle de séries chronologiques pour les erreurs d'enquête de l'équation (3.1) qui utilise les estimations directes de la variance et les autocorrélations (AC) entre les erreurs d'enquête des divers panels comme information a priori. De (3.1) il découle que les erreurs d'enquête pour la première vague sont définies comme $e_{t-j}^{t-j} = Y_{t-j}^{t-j} - \theta_{t-j}^{t-j}$. Pour les deuxième, troisième, quatrième et cinquième vagues, elles sont définies comme $e_{t-j}^{t-j} = Y_{t-j}^{t-j} - \theta_{t-j}^{t-j} - \lambda_{t-j}^{t-j} - \gamma_{t-j}^{t-j}$, pour $j = 3, 6, 9, 12$. Nous obtenons les estimations directes des variances des erreurs d'enquête pour les panels distincts au moyen de (2.2). Nous lisons ces estimations par modélisation des

3. Modèle de séries chronologiques

Les estimateurs directs, tels que l'estimateur d'Horvitz-

Thompson ou l'estimateur par régression généralisée, reposent sur l'hypothèse que le taux de chômage mensuel θ_j est un paramètre de population fixe, mais inconnu. Sous cette approche fondée sur le plan de sondage, un estimateur de θ_j pour les enquêtes transversales n'utilise que les données observées au temps t . Les données recueillies par le passé ne sont utilisées que dans le cas d'échantillons partiellement chevauchants dans un plan de sondage par panel, mais non dans le cas de plans de sondage transversaux répétés régulièrement. Scott et Smith (1974) ont proposé de considérer le paramètre de population θ_j comme une réalisation d'un processus stochastique qui peut être décrit au moyen d'un modèle de séries chronologiques.

Sous cette hypothèse, les données observées au cours des périodes précédentes $t - 1, t - 2, \dots$ peuvent être utilisées pour améliorer l'estimateur de θ_j , même dans le cas d'enquêtes dont les échantillons ne se chevauchent pas. Comme nous l'avons indiqué à la section 2.4, X'_{t-j} désigne l'estimateur par régression généralisée de θ_j basé sur le panel observé au temps t_j qui est entré dans l'enquête pour la première fois au temps $t - j$. Étant donné le schéma de renouvellement appliqué, un vecteur $X'_t = (X'_t Y'_{t-3} Y'_{t-6} Y'_{t-12})^T$ est observé chaque mois. Suivant Pfeffermann (1991), ce vecteur peut être modélisé sous la forme

$$X'_t = I_5 \pi_t + \lambda_t + \gamma_t + \epsilon_t \quad (3.1)$$

où I_5 est un vecteur pentadiagonal dont chaque élément est égal à un, $\lambda_t = (\lambda_0 \lambda_3 \lambda_6 \lambda_9 \lambda_{12})^T$ et $\gamma_t = (\gamma_0 \gamma_3 \gamma_6 \gamma_9 \gamma_{12})^T$ sont des vecteurs possédant des composantes temporelles qui tiennent compte du BGR dans la tendance et du BGR dans les composantes saisonnières respectivement, et $\epsilon_t = (\epsilon_{t-12}^T \epsilon_{t-9}^T \epsilon_{t-6}^T \epsilon_{t-3}^T \epsilon_t^T)^T$ est un vecteur dont les erreurs d'enquête correspondantes pour chaque estimation de panel. L'élaboration des modèles de séries chronologiques pour les diverses composantes de θ_j , c'est-à-dire le paramètre de population θ_j , le BGR pour la tendance λ_j , le BGR pour les composantes saisonnières γ_j , et les erreurs d'enquête ϵ_j , est décrite aux sections 3.1 à 3.3.

3.1 Modèle de séries chronologiques pour le paramètre de population

À l'aide d'un modèle structurel de séries chronologiques, nous pouvons décomposer le paramètre de population θ_j de (3.1) en une tendance, une composante saisonnière et un irrégulier, c'est-à-dire :

$$\theta_j = L_j + S_j + \epsilon_j \quad (3.2)$$

où L_j désigne une tendance stochastique, S_j une composante saisonnière stochastique et ϵ_j l'irrégulier. Pour la tendance linéaire locale, qui est défini par l'ensemble d'équations suivant :

$$\begin{aligned} L'_t &= L_{t-1} + R_{t-1} + \eta_{L,t} \\ R'_t &= R_{t-1} + \eta_{R,t} \\ E(\eta_{L,t}) &= 0, \text{Cov}(\eta_{L,t}, \eta_{L,t'}) = \begin{cases} \sigma_L^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases} \\ E(\eta_{R,t}) &= 0, \text{Cov}(\eta_{R,t}, \eta_{R,t'}) = \begin{cases} \sigma_R^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases} \end{aligned} \quad (3.3)$$

Les paramètres L_j et R_j représentent la tendance et le paramètre de pente respectivement. La composante saisonnière est modélisée sous la forme trigonométrique

$$S'_t = \sum_{l=1}^6 S_{t-l} \quad (3.4)$$

où

$$\begin{aligned} S_{t-l} &= S_{t-l-1} \cos(h_l) + S_{t-l-1}^* \sin(h_l) + \omega_{L,t}^*, l = 1, \dots, 6, \\ E(\omega_{L,t}) &= E(\omega_{L,t}^*) = 0, \\ \text{Cov}(\omega_{L,t}, \omega_{L,t'}) &= \text{Cov}(\omega_{L,t}^*, \omega_{L,t'}^*) \\ &= \begin{cases} \sigma_\omega^2 & \text{si } t = t' \text{ et } t = t' \\ 0 & \text{si } l \neq l' \text{ ou } t \neq t' \end{cases} \end{aligned} \quad (3.5)$$

La composante irrégulière ϵ_j contient la variation inexpliquée et est modélisée comme un processus de bruit blanc :

$$E(\epsilon_t) = 0, \text{Cov}(\epsilon_t, \epsilon_{t'}) = \begin{cases} \sigma_\epsilon^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases} \quad (3.6)$$

3.2 Modèle de séries chronologiques pour le biais de groupe de renouvellement

Dans l'équation (3.1), les écarts systématiques entre la

renouvellement (BGR) ; voir, par exemple, Bailliar (1975), Kumar et Lee (1983) et Pfeffermann (1991). Dans le cas de l'EPA, le BGR donne lieu à une sous-estimation systématique du niveau du taux de chômage dans les vagues d'ITAO, mais également à des écarts systématiques entre les composantes saisonnières. Le BGR est une conséquence des facteurs fortement confusionnels suivants :

- non-réponse sélective entre les vagues successives, c'est-à-dire érosion du panel ;
- différences systématiques entre les populations rejointes par les modes d'interview IPAO et ITAO. En principe, ces différences devraient être relativement faibles, puisque les numéros de téléphone sont demandés durant la première interview. Par conséquent, les numéros non publiés et les numéros de téléphone mobiles sont également appelés ;
- effets de mode de collecte, c'est-à-dire différences systématiques dans les données dues au fait que les interviews sont menées par téléphone au lieu de l'être sur place. Dans le cas de l'IPAO, la vitesse d'interview est plus lente, les répondants participent de façon plus engagée à l'interview et sont plus susceptibles de faire l'effort cognitif requis pour répondre avec soin aux questions. En outre, des réponses à caractère socialement moins désirable sont obtenues par IPAO, grâce au contact personnel avec l'interviewer. Par conséquent, moins d'erreurs de mesure sont attendues quand ce mode de collecte est utilisé (Holtbrock, Green et Krosnick 2003, et Roberts 2007). Van den Brakel (2008) décrit une expérience dans laquelle les modes de collecte par IPAO et par ITAO sont comparés à la première vague de l'EPA. Cette expérience révèle que le taux de chômage estimé est significativement plus faible dans le cas de l'ITAO ;
- fraction d'interviews par personne interposée plus grande dans le cas du mode ITAO (Van den Brakel 2008), ce qui pourrait entraîner une augmentation du nombre d'erreurs de mesure ;
- effets dus aux différences entre le questionnaire de l'IPAO et celui de l'ITAO. Ce dernier est une version fortement condensée du questionnaire de l'IPAO, car les interviews répétées portent sur les changements de situation des répondants sur le marché du travail ;
- effets de panel, c'est-à-dire les variations systématiques de comportement des membres du panel. Par exemple, les questions sur les démarches en vue de trouver un emploi posées lors de la première vague pourraient accroître les activités de recherche des chômeurs membres du panel. Les membres du panel pourraient également ajuster leurs réponses systématiquement au moment des vagues subséquentes, car ils apprennent

comment maintenir le cheminement à travers les questionnaires aussi bref que possible.

L'hypothèse est que les estimations fondées sur la première vague sont les plus fiables, puisque l'IPAO produit généralement des données de plus grande qualité et que la première vague ne souffre pas des effets de panel susmentionnés. Afin de réduire au minimum les effets du BGR, nous calons à l'heure actuelle les données des deuxième, troisième, quatrième et cinquième vagues sur celles de la première vague, comme il est décrit à la section 2.3.

2.3 Méthode d'estimation ordinaire

Les paramètres cibles concernant l'emploi et le chômage sont définis comme étant des totaux de population ou des ratios de deux totaux de population. Le taux de chômage, qui est le sujet de la présente étude, est défini comme le ratio du nombre total de chômeurs au total de la population active. Ce paramètre de population est estimé par le ratio de l'estimation par régression générale du total de la population active en chômage au total estimé de la population active. Chaque mois, les estimations de l'emploi et du chômage pour les trois mois précédents sont publiées.

Une méthode de pondération assez laborieuse est utilisée dans la méthode d'estimation ordinaire pour essayer de corriger le BGR. Nous en résumons ici les étapes les plus importantes. Pour commencer, on calcule les probabilités d'inclusion qui reflètent le plan d'échantillonnage décrit plus haut ainsi que les divers taux de réponse selon la région géographique. Puis, les poids d'inclusion pour chaque vague d'ITAO sont calés à l'aide de l'estimateur par régression généralisée sur la situation d'activité observée à la première vague. À l'étape suivante, les poids calés des vagues d'ITAO et les poids d'inclusion de la vague d'IPAO sont utilisés comme poids de sondage ou poids de départ de l'estimateur par régression généralisée, selon un schéma de pondération qui est fondé sur une combinaison de diverses classifications sociodémographiques. La méthode intégrée de pondération des personnes et des familles de Lemaitre et Dufour (1987) est appliquée pour obtenir des poids égaux pour les personnes appartenant à un même ménage. Enfin, un algorithme de détermination de bornes proposé par Huang et Fuller (1978) est appliqué pour éviter les poids négatifs. Cette méthode d'estimation est exécutée à l'aide du progiciel Bascula, Nieuwenbroek et Boonstra (2002).

Puisque cette méthode de pondération corrige à peine le BGR, une correction rigide supplémentaire est appliquée. Pour les paramètres les plus importants, le ratio entre les estimations fondées sur l'IPAO seulement et les estimations basées sur toutes les vagues de collecte est calculé en utilisant les données des 12 trimestres précédents. Les estimations pour les trois mois précédents sont multipliées par ce ratio pour les corriger du BGR.

sondage ou d'estimateurs directs, des méthodes d'estimation fondées sur un modèle doivent être utilisées pour produire des statistiques suffisamment fiables. Dans le cas d'enquêtes continues, il est possible d'appliquer un modèle structurel de séries chronologiques afin d'utiliser l'information provenant des échantillons précédents en vue d'améliorer l'exactitude des estimations. Ce modèle peut être étendu afin de tenir compte du biais de groupe de renouvellement (BGR) et de l'autocorrélation (AC) entre les divers panels de l'EPA. Cette approche, qui permet d'utiliser efficacement le plan de sondage avec renouvellement de panel de l'EPA pour estimer les chiffres mensuels concernant le marché du travail, a été proposée originellement par Pfeffermann (1991) et par Pfeffermann, Feder et Signorelli (1998). Nous appliquons ces techniques ici pour estimer le taux de chômage mensuel d'après les données de l'EPA. D'autres références à l'application de modèles de séries chronologiques en vue de produire des estimations pour des enquêtes périodiques peuvent être consultées dans Scott et Smith (1974), Scott, Smith et Jones (1977), Tam (1987), Binder et Dick (1989, 1990), Bell et Hillmer (1990), Tiller (1992), Rao et Yu (1994), Pfeffermann et Burck (1990), Pfeffermann et Rubin-Bleuer (1993), Pfeffermann et Tiller (2006), Harvey et Chung (2000), ainsi que Feder (2001).

Les estimateurs composites peuvent être considérés comme une alternative au modèle de séries chronologiques. Ils sont élaborés sous l'approche fondée sur le plan de sondage classique afin d'utiliser l'information observée aux périodes précédentes au moyen d'enquêtes périodiques avec plan de sondage avec renouvellement de panel en vue d'améliorer la précision des estimations des niveaux et des variations. Certaines références importantes concernant les estimateurs composites sont Hansen, Hurwitz et Meadow (1953), Rao et Graham (1964), Gurney et Daly (1965), Cantwell (1990), Singh (1996), Gambino, Kennedy et Singh (2001), Singh, Kennedy et Wu (2001), ainsi que Fuller et Rao (2001).

À la section 2, nous résumons le plan de sondage de l'EPA. Aux sections 3 et 4, nous élaborons un modèle structurel de séries chronologiques qui tient compte du plan de sondage avec renouvellement de panel de l'EPA. À la section 5, nous décrivons les résultats en détail. Enfin, à la section 6, nous formulons certaines remarques générales.

2. L'Enquête sur la population active des Pays-Bas

2.1 Plan d'échantillonnage

L'objectif de l'EPA des Pays-Bas est de fournir des renseignements fiables sur le marché du travail. Chaque mois, un échantillon d'adresses est sélectionné en vue d'identifier

2.2 Biais de groupe de renouvellement

Le plan avec renouvellement de panel décrit à la section 2.1 produit des différences systématiques entre les estimations du taux de chômage obtenues pour les vagues successives réalisées durant une période donnée. Dans la littérature, ce phénomène est connu sous le nom de biais de groupe de

ménages ont répondu complètement à l'enquête. Adresses en 2008. Durant cette période, environ 65 % des ensuite diminué progressivement pour atteindre 6 500 l'échantillon était, en moyenne, de 8 000 adresses. Elle a avec renouvellement de panel, la taille mensuelle brute de réinterviews. Au moment où l'EPA est passée à un plan personnes sur le marché du travail. L'interview par utilisé pour établir les changements de situation de ces Durant ces réinterviews, un questionnaire condensé est interview téléphonique assistée par ordinateur (ITAO), sont réinterviewés quatre fois à intervalle trimestriel par par personne interposée sont traités comme des ménages participer à l'enquête ne sont pas obtenues directement ou d'une ou de plusieurs personnes sélectionnées pour même ménage. Les ménages pour lesquels les réponses personne interposée est permise auprès d'autres membres du un membre sélectionné d'un ménage, l'interview par interrogés. Quand il est impossible de prendre contact avec variables cibles, seules les personnes de 15 ans et plus sont tous les membres des ménages sélectionnés. Pour les (IPAQ). Les variables démographiques sont observées pour recueillies par interview sur place assistée par ordinateur de panel. Durant la première vague, les données sont enquête continue à un plan de sondage avec renouvellement d'une EPA est passé d'une En octobre 1999, le programme de l'EPA est passé d'une 65 ans et plus sont sous-échantillonnées.

15 à 64 ans, les adresses où il n'existe que des personnes de des paramètres cibles de l'EPA ont trait à des personnes de généralement un ménage par adresse). Puisque la plupart des ménages résidant à une adresse, jusqu'à concurrence de adresses, les unités secondaires d'échantillonnage. Tous les comme étant les unités primaires d'échantillonnage, et les régions géographiques. Les municipalités sont considérées degrés stratifié d'adresses. Les strates correspondent aux fondée sur un plan d'échantillonnage en grappes à deux municipales d'enregistrement de la population. L'EPA est occupés aux Pays-Bas, qui est dressée d'après les données d'âge est la liste de toutes les adresses connues de logements établissement qui résident aux Pays-Bas. La base de son- comprend les personnes de 15 ans et plus non placées en finales d'échantillonnage. La population cible de l'EPA des ménages qui peuvent être considérés comme les unités

Estimation du taux de chômage mensuel par modélisation structurelle de séries chronologiques dans un plan de sondage avec renouvellement de panel

Jean van den Brakel et Sabine Krieg

Résumé

L'article décrit un modèle de séries chronologiques structurel multivarié qui tient compte du plan de sondage avec renouvellement de panel de l'Enquête sur la population active des Pays-Bas et qui est appliqué pour estimer les taux mensuels de chômage. Comparativement à l'estimateur par la régression généralisée, cette approche accroit considérablement la précision des estimations, grâce à la réduction de l'erreur-type et à la modélisation explicite du biais entre les vagues subséquentes de l'enquête.

Mots clés : Estimation sur petits domaines ; biais de groupe de renouvellement ; erreurs d'enquête.

1. Introduction

L'Enquête sur la population active (EPA) des Pays-Bas est fondée sur un plan de sondage avec renouvellement de panel. Chaque mois, un échantillon d'adresses est sélectionné et les données sont recueillies par interview sur place assistée par ordinateur auprès des ménages résidant à ces adresses. Ensuite, les ménages échantillonnés sont réinterviewés quatre fois par téléphone à intervalle trimestriel. La méthode d'estimation de cette enquête s'appuie sur l'estimateur par la régression généralisée élaboré par Särndal, Swensson et Wirtman (1992).

En raison des propriétés énoncées ci-après, les estimateurs par régression généralisée sont très intéressants pour la production de statistiques officielles dans un environnement de production régulier et sont par conséquent employés largement par les instituts nationaux de statistique. En premier lieu, les estimateurs par régression généralisée sont approximativement sans biais par rapport au plan, ce qui offre une forme de robustesse dans le cas d'échantillons de grande taille. Ces estimateurs sont dérivés d'un modèle de régression linéaire qui spécifie la relation entre les valeurs d'un paramètre d'intérêt et un ensemble de variables auxiliaires dont les totaux dans la population cible finie sont connus. Si le modèle de régression linéaire explicite raisonnablement bien la variation de la variable cible, il peut réduire la variance par rapport au plan ainsi que le biais dû à la non-réponse sélective (Särndal et Swensson 1987; Bethlehem 1988; Särndal et Lundström 2005). En revanche, sa spécification incorrecte pourrait accroître la variance par rapport au plan, mais les estimations ponctuelles demeurent approximativement sans biais par rapport au plan. En deuxième lieu, les estimateurs par régression généralisée sont souvent utilisés pour produire un ensemble

de poids pour l'estimation de tous les paramètres cibles d'une enquête par sondage polyvalente. Cela est non seulement commode, mais assure aussi la cohérence entre les taux marginaux de différents tableaux publiés.

Le plan avec renouvellement de panel de l'EPA et la façon dont l'estimateur par régression généralisée est appliqué dans la procédure d'estimation posent deux grands problèmes. Premièrement, il existe des différences systématiques importantes entre les vagues successives du panel dues à des effets de mode de collecte et de panel. Ce problème bien connu des plans avec renouvellement de panel est désigné dans la littérature sous le nom de biais de groupe de renouvellement (BGR); voir Baillat (1975). Dans l'EPA, le niveau du taux de chômage enregistré lors des vagues qui suivent la première est considérablement plus faible que celui observé au départ. Il existe également des différences systématiques entre les effets saisonniers pour ces vagues successives.

Deuxièmement, la taille de l'échantillon mensuel de l'EPA est trop faible pour pouvoir se fier à l'estimateur par régression généralisée pour produire des statistiques officielles mensuelles de l'emploi et du chômage. Les estimateurs par régression généralisée possèdent une variance par rapport au plan relativement importante quand la taille de l'échantillon est faible. Par conséquent, dans le cas de l'EPA, on utilise chaque mois les échantillons observés au cours des trois mois précédents pour estimer les chiffres trimestriels concernant la situation sur le marché du travail. Le principal inconvénient de cette approche est le lissage de la variation saisonnière mensuelle réelle du taux de chômage. En outre, les variations structurelles du chômage paraissent décalées dans la série de chiffres trimestriels. Puisque l'échantillon mensuel est de trop petite taille pour pouvoir se servir d'estimateurs fondés sur le plan de

- Nargundkar, M., et Joshi, G.B. (1975). Non-response in sample surveys. Dans 40th Session of the ISI, Warsaw 1975, *Contributed papers*, 626-628.
- Oh, H.L., et Scheuren, F.J. (1983). Weighing adjustments for unit non-response. Dans *Incomplete data in sample surveys*, (Eds., W.G. Madow, I. Olkin et D.B. Rubin), *Theory and bibliographies*, Academic Press, New York : Londres, 2, 143-184.
- Wand, M.P., et Jones, M.C. (1995). *Kernel Smoothing*. Londres : Chapman and Hall.
- Rust, K. (1985). Variance estimation for complex estimators in sample survey. *Journal of Official Statistics*, 381-397.
- Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41-55.

Lemme 5. Supposons que les hypothèses du théorème 1 sont vérifiées. Alors, pour tout $v \geq 1$

i) la réciproque de ϕ_i est uniformément bornée dans $i \in U_v$;

ii) les dérivées partielles de $\hat{\phi}_i^{-1}$ d'ordres un à quatre, quand elles sont évaluées à $\hat{T}_i^{-1} \hat{t}_i^{-1} = T_i^{-1} t_i^{-1} = \delta_i = 0$ et $\delta_2 = 0$, sont uniformément bornées en $i \in U_v$;

iii) $E(\hat{\phi}_i^{-1})$ est uniformément bornée en $i \in U_v$;

iv) l'inverse de $\hat{\phi}_i$ satisfait

$$\hat{\phi}_i^{-1} = \hat{\phi}_i^{-1} - \hat{\phi}_i^{-1} e_i' T_i^{-1} (\hat{t}_i^{-1} - \hat{T}_i^{-1} B_i) + \varepsilon_{iv} + O\left(\frac{N_i^v h_i^2}{1}\right), \tag{22}$$

uniformément en $i \in U_v$, où les ε_{iv} sont des variables aléatoires telles que

$$\max_{i \in U_v} E\left(\varepsilon_{iv}^2\right) = O\left(\frac{n_i^2 h_i^2}{1}\right).$$

Lemme 6. Supposons que les hypothèses du théorème 1 sont vérifiées. Définissons les variables aléatoires $\bar{y}^{\pi\phi_v}, \bar{d}^{\pi\phi_v}$ et $\bar{y}^{\pi\phi_v}, \bar{d}^{\pi\phi_v}$ par

$$(\bar{y}^{\pi\phi_v}, \bar{d}^{\pi\phi_v}, \bar{\varepsilon}^{\pi\phi_v})' = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} (1, \phi_i^{-1} e_i' T_i^{-1} (\hat{t}_i^{-1} - \hat{T}_i^{-1} B_i), \varepsilon_{iv})' y_i^1 R_i.$$

Alors,

$$E(\bar{y}^{\pi\phi_v} - \bar{y}^{\pi N_v}) = \begin{cases} O(h_{k+1/2}^v) & k \text{ pair,} \\ O(h_{k-1}^v) & k \text{ impair,} \end{cases} \tag{23}$$

$$\text{Var}(\bar{y}^{\pi\phi_v}) = O\left(\frac{n_v}{1}\right), \tag{24}$$

$$E[\bar{d}^{\pi\phi_v} | E[\bar{d}^{\pi\phi_v} | A]] = O\left(\frac{n_v h_v}{1}\right) \tag{25}$$

et

$$E(\varepsilon_{iv}^2) = O\left(\frac{n_i^2 h_i^2}{1}\right). \tag{26}$$

Bibliographie

Alho, J.M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika*, 617-624.

Berger, Y.G., et Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 67(1), 79-89.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 1026-1053.

Da Silva, D.N., et Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics*, 4, 563-579.

Da Silva, D.N., et Opsomer, J.D. (2008). Theoretical properties of propensity weighting for survey nonresponse through local polynomial regression. Rapport technique #2008/6, Department of Statistics, Colorado State University.

David, M.H., Little, R., Samuhel, M. et Tiesi, R. (1983). Imputation models based on the propensity to respond. Dans *ASA Proceedings of the Business and Economic Statistics Section*, 168-173.

Eklöhm, A., et Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 325-337.

Fan, J., et Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*. Londres : Chapman & Hall.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. Dans *ASA Proceedings of the Social Statistics Section*, 197-202.

Gioinni, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron*, 4, 185-200.

Groves, R., Dillman, D., Eitinge, J. et Little, R.J.A. (2002). *Survey Nonresponse*. New York : John Wiley & Sons, Inc.

Hastie, T.J., et Tibshirani, R.J. (1986). Generalized additive models. *Statistical Science*, 297-318.

Iannacchione, V.G., Milne, J.G. et Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. Dans *ASA Proceedings of the Section on Survey Research Methods*, 637-642.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 89-96.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 4, 501-514.

Laaksonen, S. (2006). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications*, 2, 95-100.

Loeh, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove : Duxbury Press.

(C5) la dérivée première $f_x^X(\cdot)$ est continuellement différentiable et contient un nombre fini de changements de signe sur $\text{supp}(f_X)$. La dérivée première $K'(\cdot)$ possède un nombre fini de changements de signe sur $\text{supp}(K)$;

(C6) la matrice $N_\nu \mathbf{T}^{-1}$ est non singulière pour tout $i \in U_\nu$ et tout $\nu \geq 1$.

A.2 Calculs techniques

Les preuves complètes sont données dans Da Silva et Opsomer (2008). La preuve du théorème 1 s'appuie sur la

détermination de bornes pour les moments de la différence $\bar{y}^{\text{np}\nu} - \bar{y}^{\text{pq}\nu}$ sous le mécanisme probabiliste du plan de sondage et du modèle de réponse combinés, suivie par le calcul des taux de convergence pour le biais et la variance de l'estimateur linéarisé $\bar{y}^{\text{np}\nu}$. Cela est accompli au moyen d'une série de six lemmes, qui sont énoncés ici sans preuve. La preuve du théorème 2 est fondée sur le résultat du théorème 1, suivie par une linéarisation supplémentaire de la forme ratio.

Dans la suite, pour simplifier la notation, nous sup-primerons le fait que les résultats sont conditionnés sur les séries $\mathbf{x}_\nu = (x_1, \dots, x_{N_\nu})$ dans les populations U_ν . Cependant, nous avons montré dans Da Silva et Opsomer (2008), comme nous l'avions fait dans Da Silva et Opsomer (2006), que les résultats que contiennent ces lemmes sont vérifiés avec une probabilité de 1 sur ces séries. Donc, les résultats peuvent être interprétés comme vérifiés pour toutes les séries dans les populations, sauf sur un ensemble de probabilités 0 par rapport à la distribution des \mathbf{x}_ν .

Lemme 1. Supposons que les hypothèses (C1) à (C5) sont vérifiées. Considérons $\mu_f(K, x) = \int_0^1 z' K(z) dz$, où $D_{x, h_\nu} = \{t : (x + ht) \in \text{supp}(f_X)\} \cap \text{supp}(K)$. Alors, pour tout $\ell = 0, 1, \dots, k + 2$,

$$\sup_{\nu \rightarrow \infty} \left| \frac{N_\nu h_\nu^{(f_X)}}{1} \sum_{j \in U_\nu} K \left(\frac{h_\nu}{X_j - x} \right) (X_j - x)^\ell - E_\nu(x, \ell) \right| \rightarrow 0,$$

où

$$E_\nu(x, \ell) = f_X^X(x) \mu_\ell(K, x) h_\nu^\ell + f_X^X(x) \mu_{\ell+1}(K, x) h_{\nu+1}^\ell + o(h_{\nu+1}^\ell).$$

Lemme 2. Supposons que les hypothèses (C1) à (C5) sont vérifiées. Considérons l'ajustement à la population $\phi_i = \phi(x_i, k, h_\nu)$, $i \in U_\nu$, défini en (10). Donc, pour tout $i \in U_\nu$, il existe des termes bornés positifs $c_1(x_i)$, $c_2(x_i)$ et $c_3(x_i)$, tels que, si x_i dans un point intérieur de $\text{supp}(f_X)$

$$\phi_i - \phi(x_i) = \begin{cases} c_1(x_i) h_{k+2}^\nu + o(h_{k+2}^\nu) & k \text{ est pair} \\ c_2(x_i) h_{k+1}^\nu + o(h_{k+1}^\nu) & k \text{ est impair} \end{cases}$$

et si x_i dans un point au bord de $\text{supp}(f_X)$

$$\phi_i - \phi(x_i) = c_3(x_i) h_{k+1}^\nu + o(h_{k+1}^\nu),$$

où tous les termes d'ordre plus faible tiennent uniformément dans $i \in U_\nu$.

Lemme 3. Supposons que les hypothèses (C1) et (C4) sont vérifiées. Alors,

i) pour $p \in [0, \infty)$ fixe,

$$\limsup_{\nu \rightarrow \infty} \frac{1}{2N_\nu h_\nu} \sum_{j \in U_\nu} I_{\{x_j \in [0, h_\nu] \cap (1-h_\nu, 1]\}} > \infty;$$

$$\limsup_{\nu \rightarrow \infty} \frac{N_\nu}{1} \sum_{j \in U_\nu} I_{\{x_j \in (h_\nu, 1-h_\nu)\}} > \infty;$$

ii) il existe ν^* , indépendant de x , tel que, quand $\nu \geq \nu^*$,

$$\sum_{j \in U_\nu} I_{\{|x_j - x| \leq h_\nu\}} \geq k + 1.$$

Lemme 4. Supposons que les hypothèses du théorème 1 sont vérifiées. Considérons les matrices $\mathbf{T}_{sl, pd}^{\nu} = \{\mathbf{T}_{sl, pd}^{\nu}\}$ et $\mathbf{T}_{sl, p}^{\nu} = \{\mathbf{T}_{sl, p}^{\nu}\}$ et les vecteurs $\mathbf{t}_{sl, p}^{\nu} = \{t_{sl, p}^{\nu}\}$ et $\mathbf{t}_{sl, p}^{\nu} = \{t_{sl, p}^{\nu}\}$ donnés dans (7) et (10). Alors,

- les $N_\nu^{-1} \mathbf{T}_{sl, pd}^{\nu}$ et $N_\nu^{-1} \mathbf{t}_{sl, p}^{\nu}$ sont uniformément bornés dans $i \in U_\nu$, pour tout $p, q = 1, \dots, k + 1$;
- les $\mathbf{T}_{sl, pd}^{\nu}$ et $\mathbf{t}_{sl, p}^{\nu}$ satisfont

$$\max_{1 \leq p, q \leq k+1} \mathbf{E} \left(\frac{N_\nu}{\mathbf{T}_{sl, pd}^{\nu}} - \mathbf{T}_{sl, pd}^{\nu} \right) = O \left(\frac{N_\nu}{\mathbf{T}_{sl, pd}^{\nu}} \right)$$

$$\max_{1 \leq p \leq k+1} \mathbf{E} \left(\frac{N_\nu}{\mathbf{t}_{sl, p}^{\nu}} - \mathbf{t}_{sl, p}^{\nu} \right) = O \left(\frac{N_\nu}{\mathbf{t}_{sl, p}^{\nu}} \right)$$

uniformément dans $i \in U_\nu$;

iii) la variable aléatoire $\mathbf{e}_i' \mathbf{T}_{sl, p}^{\nu} \mathbf{t}_{sl, p}^{\nu}$ satisfait

$$\max_{i \in U_\nu} \mathbf{E} \left(\mathbf{e}_i' \mathbf{T}_{sl, p}^{\nu} \mathbf{t}_{sl, p}^{\nu} \right) = O \left(\frac{N_\nu}{1} \right)$$

$$\max_{i \in U_\nu} \mathbf{E} \left(\mathbf{e}_i' \mathbf{T}_{sl, p}^{\nu} \mathbf{t}_{sl, p}^{\nu} \right) = O \left(\frac{N_\nu}{1} \right).$$

Un certain nombre de questions concernant l'application de la méthode dans les enquêtes réelles, que ce soit dans le cas unitaire décrit en détail ici ou dans les diverses extensions du modèle que nous venons de mentionner, doivent être étudiées plus en profondeur. Un important problème pratique est celui du choix du paramétrage de l'estimateur, tel que le degré du polynôme local et la largeur de la fenêtre. Comme il est mentionné dans la littérature non paramétrique (par exemple, Fan et Gijbels 1996, page 77) et confirmé dans les simulations, l'utilisation de polynômes de degré plus élevé réduit le biais mais accroît la variance, de sorte que les polynômes de degré $k = 1$ ou 2 sont généralement considérés comme un bon compromis. Le choix de la largeur de fenêtre est plus critique. Dans nos simulations, les résultats n'étaient que moyennement sensibles à ce choix dans une fourchette « raisonnable » de valeur, c'est-à-dire celles qui garantissent que le nombre d'observations utilisées pour estimer $\phi(x)$ à tout x ne devienne pas trop petit (voir la discussion à la fin de la section 2) ou si grand que l'ajustement du modèle ne permette pas de saisir les variations dans $\phi(\cdot)$ sur l'intervalle de valeur de x . À titre de règle empirique, nous recommandons de considérer pour h des valeurs comprises entre 20 % et 50 % de l'étendue de x comme bon point de départ, puis de procéder à une détermination finale en examinant à la fois les diagnostics de modélisation concernant l'ajustement du modèle $\phi(x)$ et les diagnostics de pondération concernant les poids de sondage corrigés $(\pi_i \phi_i)^{-1}$, comme cela se ferait pour construire des poids dans les cellules.

Remerciements

Nous remercions le rédacteur associé et deux examinateurs de leurs commentaires constructifs. Les travaux du premier auteur ont été financés par le CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) du Brésil aux termes de la subvention Projeto Universal 480518/2004-1.

Annexe

A.1 Hypothèses

Nous énonçons maintenant les hypothèses nécessaires pour obtenir nos principaux résultats. Une discussion détaillée de ces hypothèses figure dans Da Silva et Opsomer (2008). Considérons le cadre asymptotique de la section 3. Soit $I_v = (I_1, I_2, \dots, I_{N_v})'$ le vecteur d'indicateurs d'inclusion dans l'échantillon pour la v^e population. En supprimant le v pour simplifier la notation, posons $\pi_i = \Pr(I_i = 1)$, et désignons par

$$\Delta^{j_1, \dots, j_k} \equiv E_d \left(\prod_{i=1}^k (U_{j_i} - \pi_{j_i}) \right) \quad (19)$$

les moments de degré plus élevé pour les indicateurs d'inclusion dans l'échantillon $I_{j_1}, I_{j_2}, \dots, I_{j_k}$ par rapport au plan d'échantillonnage. Nous supposons qu'il existe des constantes positives $\lambda_1, \lambda_2, \dots, \lambda_6$ telles que :

(A1) $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \lambda_5 < \lambda_6$, $\forall i \in U_v$;

(A2) $N_v^{-1} \pi_i \rightarrow \pi$, pour $0 < \pi < 1$, quand $v \rightarrow \infty$;

(A3) Pour des $j_1, j_2, \dots, j_k \in U_v$, distincts, où $k = 2, 3, \dots, 8$,

$$\left| \Delta^{j_1, \dots, j_k} \right| \leq \begin{cases} \prod_{i=1}^k (N - \ell + 1)^{-1} \lambda_i^k & \text{si } k \text{ est pair,} \\ \prod_{i=1}^k (N - \ell + 1)^{-1} \lambda_i^{k-1} \lambda_4 & \text{si } k \text{ est impair} \end{cases}$$

(A4) $\lim_{v \rightarrow \infty} N_v^{-1} \sum_{i \in U_v} \lambda_i^4 = 1$ et $N_v^{-1} \sum_{i \in U_v} \lambda_i^4 \leq \lambda_5$, pour tout $v \geq 1$.

Soit $R_v = (R_1, R_2, \dots, R_{N_v})'$ le vecteur d'indicateurs de réponse pour la v^e population. En plus des hypothèses concernant le plan d'échantillonnage et la distribution de la population de la variable Y , nous avons besoin des hypothèses suivantes concernant le mécanisme de réponse :

(B1) R_1, R_2, \dots, R_{N_v} sont des variables aléatoires indépendantes ;

(B2) $\Pr\{R_i = 1 \mid I_v, Y_v, x_v\} = \Pr\{R_i = 1 \mid x_v\}, \forall i \in U_v$;

(B3) $\phi_i = \phi(x_i), \forall i \in U_v$, où $\phi(\cdot)$ est une $(k+2)^e$ fonction continuellement différentiable avec $\lambda_6 < \phi(\cdot) \leq 1$. La dérivée première $\phi(\cdot)$ possède un nombre fini de changements de signe.

En ce qui concerne la distribution des x_i et l'estimateur par noyau, nous supposons que :

(C1) pour tout $v \geq 1$, x_1, x_2, \dots, x_{N_v} sont les réalisations des variables aléatoires X_1, X_2, \dots, X_{N_v} indépendantes et identiquement distribuées suivant la loi $F_X(x) = \int_{-\infty}^x f_X(t) dt$, où $f_X(\cdot)$ est une densité de probabilité continue et positive sur un ensemble compact $[a_X, b_X]$;

(C2) la fonction noyau $K(\cdot)$ est une densité de probabilité bornée et continue, qui est symétrique autour de zéro et appartient à l'intervalle $[-1, 1]$;

(C3) $\int |z|^{k+4} K(z) dz < \infty$;

(C4) pour tout $v \geq 1$, $\{h_v\}$ est une série de largeurs de fenêtre satisfaisant $0 < h_v \leq 1$, $h_v \rightarrow 0$, $n_v h_v^2 \rightarrow \infty$, quand $v \rightarrow \infty$;

cas pour toutes les variables. Comme il est discuté dans Da Silva et Opsomer (2006), les méthodes de rééchantillonnage pour les estimateurs corrigés de la non-réponse offrent souvent de tenir compte d'une composante de la variance totale qui inclut l'effet de l'échantillonnage ainsi que du mécanisme de réponse. Nous concluons, par conséquent, que les propriétés de biais différentes manifestées pour les diverses variables pourraient être dues à cette composante manquante de la variance.

6. Conclusion

Dans le présent article, nous avons étudié les propriétés de la pondération non paramétrique par la propension à répondre en tant que méthode de correction de la non-réponse aux enquêtes. La régression par polynômes locaux semble offrir un moyen flexible de production de nouveaux ajustements pour tenir compte de la non-réponse. Les résultats présentés étendent ceux décrits dans Da Silva et Opsomer (2006) en permettant l'utilisation de polynômes locaux de degré arbitraire, ce qui offre des avantages aussi bien théoriques que pratiques par rapport à la régression par noyau de degré nul.

L'expérience par simulation a révélé qu'outre ses bonnes propriétés théoriques, l'estimateur donne des résultats comparables à un estimateur fondé sur un modèle paramétrique spécifié correctement en ce qui concerne le biais et la variance, tout en offrant une protection contre une éventuelle erreur de spécification du modèle. L'estimateur

À la section 5, nous avons appliqué la correction non paramétrique de la non-réponse aux données de la NHANES en modélisant la probabilité de réponse sous forme d'une fonction lisse de l'âge des répondants et en pondérant les données par l'inverse des probabilités de réponse estimées. La même approche peut être appliquée à d'autres ensembles de données d'enquête quand des covariables continues reliées à la probabilité de réponse sont disponibles pour toutes les unités figurant dans l'échantillon original. Il s'agit d'une alternative viable à l'approche de pondération à l'intérieur de la cellule adoptée couramment dans des situations où les cellules sont constituées par « binning », ou groupement par classe, sur une ou plusieurs variables continues.

Fonction de propension logistiqu						Fonction de propension logistiqu					
(prédicteur non linéaire)						(prédicteur linéaire)					
PAD						PAD					
HYP						HYP					
CTE						CTE					
PAS						PAS					
Propensions réelles à répondre						Régession pondérée par polynômes locaux :					
Type de correction						Régession pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré 3					
CTE						Régession non pondérée par polynômes locaux :					
PAS						Degré 0					
PAD						Degré 1					
HYP						Degré 2					
CTE						Degré 3					
PAS						Régession non pondérée par polynômes locaux :					
PAD						Degré 0					
HYP						Degré 1					
CTE						Degré 2					
PAS						Degré 3					
PAD						Régession non pondérée par polynômes locaux :					
HYP						Degré 0					
CTE						Degré 1					
PAS						Degré 2					
PAD						Degré 3					
HYP						Régession non pondérée par polynômes locaux :					
CTE						Degré 0					
PAS						Degré 1					
PAD						Degré 2					
HYP						Degré					

Tableau 2
Variances Monte Carlo normalisées des estimateurs corrigés de la non-réponse de la pression artérielle systolique (PAS) moyenne, la pression artérielle diastolique (PAD), l'indicateur d'hypertension (HYP) et l'indicateur de cholestérol total sérique élevé (CTE) basés sur 1 000 ensembles de réponses pour deux fonctions de propension à répondre selon l'âge du participant à l'enquête dans la NHANES de 2005-2006

Type de correction		Fonction de propension logistiqu						Fonction de propension logistiqu					
		(prédicteur linéaire)			(prédicteur non linéaire)			(prédicteur linéaire)			(prédicteur non linéaire)		
		PAS	PAD	HYP	CTE	PAS	PAD	HYP	CTE	PAS	PAD	HYP	CTE
Propositions réelles à répondre		100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Régression logistiqu		85,9	92,4	79,5	96,5	63,9	61,5	54,1	52,0	67,4	64,4	54,1	52,0
Régression pondérée par polynômes locaux :		74,9	81,2	65,7	92,1	70,3	67,0	67,4	75,0	75,0	76,0	96,1	75,0
Degré 0		74,9	81,2	65,7	92,1	70,3	67,0	67,4	75,0	75,0	76,0	96,1	75,0
Degré 1		81,8	89,5	66,2	92,7	73,6	69,8	68,9	76,0	76,0	76,0	96,1	76,0
Degré 2		81,3	89,8	65,5	94,0	90,3	81,7	88,0	96,1	96,1	96,1	96,1	96,1
Degré 3		82,3	90,2	65,8	93,1	90,1	82,2	87,7	96,2	96,2	96,2	96,2	96,2
Régression non pondérée par polynômes locaux :		82,2	85,8	77,6	95,8	71,9	69,2	70,7	74,7	74,7	74,7	74,7	74,7
Degré 0		82,2	85,8	77,6	95,8	71,9	69,2	70,7	74,7	74,7	74,7	74,7	74,7
Degré 1		83,6	90,1	79,4	95,7	74,4	71,1	71,2	74,6	74,6	74,6	74,6	74,6
Degré 2		86,6	91,3	79,3	96,1	91,8	84,5	91,8	96,8	96,8	96,8	96,8	96,8
Degré 3		87,3	91,5	78,5	95,0	91,2	84,7	91,2	96,9	96,9	96,9	96,9	96,9
Pondération à l'intérieur de la cellule		79,7	89,1	62,1	91,6	82,5	77,0	81,1	92,3	92,3	92,3	92,3	92,3
Naïve		71,3	58,0	81,7	74,6	48,6	48,7	45,5	45,1	45,1	45,1	45,1	45,1

$$w_i^{(b)} = \begin{cases} 0, & \text{pour un participant à l'enquête} \\ & i \in \text{UPF}, j, j \in s_i \\ & i \in \text{UPF}, j, j \in s_i \\ n_i(n_i - 1)w_i, & \text{pour un participant à l'enquête} \\ & i \in \text{UPF}, j, j \in s_i, (j' \neq j) \\ w_i, & \text{pour un participant à l'enquête} \\ & i \notin s_i, \end{cases}$$

Nous avons également appliqué ces poids dans l'estimation des propensions à répondre pour la procédure de correction par la régression par polynômes locaux pondérée.

Nous avons appliqué l'estimateur jackknife de variance (18) à chaque vecteur de réponses provenant des deux fonctions de propension, ce qui a donné les estimations $\hat{V}^{JK}(\hat{\theta}(b))$, $b = 1, 2, \dots, B$, pour tous les estimateurs corrigés dans l'expérience. Aux fins de comparaison, il serait intéressant de produire des estimations des variances correspondantes par la méthode Monte Carlo. Cependant, comme l'échantillon de la NHANES est fixe, la variance Monte Carlo des estimations ponctuelles $\theta(b)$ sur les vecteurs de réponses estime uniquement la variance conditionnelle $\text{Var}(\theta|s_y)$ par rapport au modèle de réponse.

$$\text{Var}(\theta) = \text{Var}(E(\theta|s_y)) + E(\text{Var}(\theta|s_y)),$$

où les moments « internes » sont calculés par rapport au modèle de réponse sachant que l'échantillon et les moments « externes » sont calculés par rapport au plan d'échantillonnage, la variance sous le plan de $E(\theta|s_y)$ doit être

$$\hat{V}^C(\theta) = \hat{V}^{JK}(\bar{Y}^{\pi_{\text{rai}}}) + \frac{B-1}{B} \sum_{b=1}^B (\hat{\theta}(b) - \bar{\theta})^2.$$

prise en compte afin d'obtenir une cible d'estimation valide pour $\hat{V}^{JK}(\theta)$. Étant donné que la régression pondérée et non pondérée par polynômes locaux et la pondération à l'intérieur de la cellule produisent des estimateurs approximatifs conditionnellement sans biais de l'estimateur sur échantillon complet, $\bar{Y}^{\pi_{\text{rai}}} = \sum_{i \in s_y} w_i Y_i^j / \sum_{i \in s_y} w_i$, pour les deux fonctions de propension à répondre, nous avons décidé d'utiliser l'estimateur jackknife de la variance de $\bar{Y}^{\pi_{\text{rai}}}$ comme « substitut » de $\text{Var}(E(\theta|s_y))$. Donc, notre « variance de comparaison » sera définie sous la forme

L'utilisation de $\hat{V}^C(\theta)$ au lieu de la variance réelle aura tendance à sous-estimer tout problème de biais associé à l'utilisation de l'estimateur jackknife de la variance pour l'estimateur sur échantillon complet. Cependant, elle montrera dans quelle mesure la méthode de rééchantillonnage réussit bien à refléter la variabilité due à la non-réponse. Le tableau 3 donne les biais relatifs des estimations jackknife de la variance obtenus dans l'expérience. Les résultats montrent que l'estimateur jackknife de la variance concorde mieux avec la variance de comparaison que la version non pondérée de cette méthode. Les résultats pour la fonction à prédicteur non linéaire révèlent un biais plus important que ceux obtenus pour le prédicteur linéaire, les biais positifs et négatifs étant plus prononcés dans le premier

Parmi les estimateurs affectés par la non-réponse produite, les résultats les plus biaisés sont clairement ceux obtenus pour l'estimateur « naïf » non corrigé. Comme le montre la dernière ligne du tableau 1, le biais est élevé pour les estimations de la prévalence de l'hypertension et de la pression artérielle systolique moyenne, car ces caractéristiques (variables) étudiées sont celles dont la corrélation avec la variable AGF est la plus forte, ainsi que pour l'estimation de la prévalence du cholestérol total sérique élevé. Les biais de l'estimateur naïf peuvent être réduits avec succès au moyen de l'estimateur fondé sur les propensions réelles à répondre, n'importe lequel des estimateurs corrigés par la régression par polynômes locaux, l'estimateur avec pondération à l'intérieur de la cellule ou l'estimateur corrigé par régression logistique, si le modèle de la fonction de propension est spécifié correctement. En ce qui concerne la réduction du biais, les meilleurs résultats sont obtenus en utilisant l'estimateur corrigé par les propensions réelles à répondre, parce qu'il est conditionnellement sans biais pour les estimations en échantillon complet. L'ajustement logistique, s'il est appliqué sous le modèle correct, donne par la fonction de propension à répondre avec un prédicteur linéaire, produit aussi des estimations presque sans biais. Pour la deuxième fonction de propension à répondre, où la forme du prédicteur n'est pas bien reflétée par l'ajustement d'une courbe de régression logistique, cette correction donne un estimateur conditionnellement biaisé.

Les moyennes des estimations par la régression par polynômes locaux se rapprochent généralement des estimations en échantillon complet si l'on augmente le degré du polynôme ajusté, le rapprochement le plus important étant obtenu en passant d'une constante locale à un estimateur linéaire local. Donc, il semble que la régression par polynômes locaux est effectivement supérieure à la régression par la méthode du noyau dans ce contexte. La différence entre les formes pondérées et non pondérées de cet ajustement est très faible et les deux méthodes produisent des biais conditionnels plus petits dans l'ensemble que les biais de l'estimateur avec pondération à l'intérieur de la cellule, quand elles sont mises en œuvre par ajustement local d'un polynôme d'ordre supérieur à zéro pour estimer les propensions à répondre. Les estimateurs pondérés et non pondérés corrigés par la propension à répondre en ajustant un polynôme de degré nul ont un biais plus petit aux petites largeurs de fenêtre, comme nous l'avons observé pour la largeur de fenêtre 0,15, par exemple, mais les petites largeurs de fenêtre ont tendance à accroître la variance des estimateurs. Dans l'ensemble, les ajustements par la régression pondérée et non pondérée par polynômes locaux donnent de meilleurs résultats que l'ajustement par régression logistique paramétrique quand le modèle de réponse est spécifié incorrectement. L'application des corrections par la régression

par polynômes locaux de degré supérieur à un donne des résultats semblables à ceux de l'ajustement logistique sous spécification correcte du modèle de réponse.

5.4 Variance et estimation de la variance

Le tableau 2 donne les variances des méthodes d'ajustement étudiées ici sur l'ensemble des répliques de non-réponses, normalisées, par souci de clarté, par la variance de l'ajustement par la propension à répondre réelle. Fait intéressant, il semble exister une relation inverse entre la grandeur des biais relatifs présentés au tableau 1 et la variance présente dans ce tableau. Dans les cas où le biais relatif est faible (régression pondérée et non pondérée par polynômes locaux, pondération à l'intérieur de la cellule, ainsi qu'ajustement par régression logistique pour la fonction de propension linéaire), toutes les méthodes semblent mener à des variances à peu près similaires. Les polynômes locaux de degré élevé ont tendance à être plus variables que ceux de degré plus faible, et ce particulièrement dans le cas de la fonction de propension non linéaire, pour laquelle un saut net est observé lorsque l'on passe du degré 1 (local linéaire) au degré 2 (local quadratique). Dans l'ensemble, la régression linéaire locale, pondérée ou non pondérée, semble offrir un bon compromis entre le biais et la variance de la méthode de correction de la non-réponse.

Les résultats de simulation susmentionnés montrent le comportement de plusieurs ajustements pour corriger la non-réponse dans les conditions de la NHANES. Nous allons maintenant examiner l'approche d'estimation de la variance par rééchantillonnage de la section 4 et évaluer son utilité en tant que mesure de l'incertitude fondée sur l'échantillon pour les estimateurs corrigés de la non-réponse dans les mêmes conditions. Nous avons appliqué l'estimateur (17) avec la méthode du jackknife. Comme la NHANES ne fournit pas d'information sur les probabilités conjointes d'inclusion dans l'échantillon, nous n'avons pas pu nous servir d'un estimateur de variance jackknife complet comme, par exemple, dans Berger et Skinner (2005), afin de tenir compte de la sélection des unités avec probabilité variable dans l'enquête. Par conséquent, nous avons supposé que les plans d'échantillonnage à l'intérieur des strates de la NHANES pouvaient être approximatés par l'échantillonnage en grappes avec remise et avons réécrit (17) sous la forme proposée par Rust (1985),

$$\hat{V}_{JK}(\theta) = \sum_{j \in s_j} c_j^2 \sum_{i \in s_i} (\theta^{(j)} - \theta)^2, \quad (18)$$

où s_j désigne l'ensemble d'unités dans l'échantillon qui proviennent de la j^{e} strate de la NHANES, $i = 1, 2, \dots, T, n_j$ est le nombre d'unités sélectionnées dans s_j , $c_j = (n_j - 1)/n_j$ et $\theta^{(j)}$ est obtenu à partir de (5) en remplaçant les w_i par les poids de rééchantillonnage

quartiles d'échantillon de cette variable. L'échantillon a été subdivisé ainsi en 60 cellules. Soit s_{ij}^x l'ensemble d'éléments échantillonnés et $1en_{ij}^x$ l'ensemble d'éléments répondants, respectivement, dans la i^e cellule. Alors, la correction par pondération dans la cellule est défini en prenant

6. Correction naïve : $\hat{\phi}_i = 1, i \in s_v$ pour tous les répondants $i \in s_{rg}$.

$$\sum_{i \in S^k} \frac{m_i}{\sum_{i \in S^k} m_i} = \phi$$

ais et robuste à l'erreur de spécification de la fonction de propension à répondre

échantillon complet sans non-réponse artificielle est les moyennes estimées de Hajek pour les quatre études sont respectivement PAS=122,19 mm Hg, HXP=39,04 % et STE=15,76 %.

3. Régression pondérée par polynômes locaux de degré k et de largeur de fenêtre h : $\hat{\phi} = \hat{\phi}(x, k, h)$ donnée par (8), avec $t \in s$, $k = 0, 1, 2, 3$, $h = 0, 15, 0, 25, 0, 50$ et la fonction moyen d'Esparechnikov $K(t) = (3/4)(1 - x^2)I|x| \leq 1$.

- évaluer l'effet du passage d'une constante locale à des polynômes d'ordre plus élevé.

Tableau 1
Biais relatifs (%) des estimateurs corrigés de la non-réponse de la pression artérielle systolique (PAS) moyenne, la pression artérielle diastolique (PAD), l'indicateur d'hypertension (IH) et l'indicateur de cholestérol total sérique élevé (CTE), basés sur 1 000 ensembles de réponses pour deux fonctions de propension à répondre selon l'âge du participant à l'enquête dans la NHANES de 2005-2006

[illegible]

5. Application aux données de la NHANES

5.1 Le plan de sondage de la NHANES

Nous évaluons la performance des estimateurs corrigés par la régression par polynômes locaux au moyen de données réelles. Nous utilisons pour cela les données de la vague de 2005-2006 de la National Health and Nutrition Examination Survey (NHANES) menée par le National Center for Health Statistics des Centers for Disease Control and Prevention (NCHS/CDC) du U.S. Department of Health and Human Services. Cette enquête est réalisée auprès d'un échantillon à plusieurs degrés, stratifié, de la population civile américaine non placée en établissement. Dans les grandes lignes, la formation de l'échantillon est la suivante :

- i) dans chaque strate, les unités primaires d'échantillonnage (UPE) correspondent à des comtés ou à des groupements de petits comtés qui sont sélectionnés par échantillonnage avec probabilité proportionnelle à une mesure de taille ;
- ii) parmi les UPE échantillonnées, des groupes d'ilots urbains (segments) contenant des grappes de ménages sont sélectionnés, également par échantillonnage avec probabilité proportionnelle à la taille ;
- iii) dans les segments sélectionnés, les grappes de ménages sont sélectionnées aléatoirement avec probabilité de sélection variable afin de suréchantillonner certains groupes d'âge, d'origine ethnique ou de revenu dans certaines régions géographiques ;
- iv) dans les ménages sélectionnés, un ou plusieurs participants sont sélectionnés aléatoirement.

La diffusion publique des données de la NHANES comporte deux aspects importants. Premièrement, afin de réduire les risques de divulgation, le plan d'échantillonnage à quatre degrés, stratifié est condensé en un plan à un degré stratifié dans lequel ni la nouvelle variable de strate ni la nouvelle variable d'UPE ne correspond à la même variable dans le plan de sondage original. Deuxièmement, les poids d'échantillonnage de base, qui sont égaux à l'inverse des probabilités d'inclusion des participants à l'enquête, ne sont pas diffusés. Les poids fournis reflètent les ajustements apportés aux poids de base pour tenir compte de la non-réponse totale dans les volets interview et examen physique de l'enquête, et pour produire des estimations qui concordent avec les totaux de contrôle connus de population.

5.2 L'expérience par simulation

Afin d'évaluer empiriquement les estimateurs de régression par polynômes locaux en tant que correction pour la non-réponse dans les enquêtes complexes, nous allons appliquer une source artificielle de non-réponse totale à

où les coefficients de régression β_0, \dots, β_j sont choisis de sorte que les fonctions de propension à répondre donnent un taux de non-réponse global d'environ 30 % quand elles sont appliquées aux valeurs d'échantillon de x . Dans les deux cas, nous avons maintenu l'échantillon de la NHANES fixe et produit $B = 1\,000$ vecteurs indépendants d'indicateurs de réponse par échantillonnage de Poisson.

Predicteur non linéaire :

$$\phi_j(x) = \{1 + \exp[-(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \cos(\beta_4 x^2/\pi) \sin(\beta_5 x/\pi))]\}^{-1}$$

Predicteur linéaire :

$$\phi_j(x) = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1}$$

non linéaire de x de la forme suivante

Predicteur linéaire :

$\phi_j(x) = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1}$

Predicteur non linéaire :

$\phi_j(x) = \{1 + \exp[-(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \cos(\beta_4 x^2/\pi) \sin(\beta_5 x/\pi))]\}^{-1}$

L'ensemble de données à grande diffusion de la NHANES. Nous considérons que le mécanisme de non-réponse est une fonction lisse de l'âge, exprimé en années, des participants à l'enquête (AGE). Pour cette comparaison, nous choisissons comme variables étudiées quatre caractéristiques associées aux maladies cardiaques, à savoir la pression artérielle systolique (PAS), la pression artérielle diastolique (PAD), l'indicateur d'hypertension (HYP) et l'indicateur de taux de cholestérol total sérique élevé (CTE). Toutes ces variables ont été mesurées chez les participants à l'enquête ayant 18 ans ou plus. Pour les variables de pression systolique et diastolique, les valeurs ont été obtenues en calculant la moyenne, pour les mesures correspondantes, d'un ensemble comptant jusqu'à quatre lectures. Nous avons considéré comme faisant de l'hypertension les individus dont la pression artérielle systolique était égale ou supérieure à 140 mm Hg ou dont la pression artérielle diastolique moyenne était égale ou supérieure à 90 mm Hg ou qui prenaient, au moment de l'enquête, un médicament pour faire baisser la tension. Nous avons considéré comme ayant un taux de cholestérol total sérique élevé les personnes dont le taux de cholestérol sérique total était égal ou supérieur à 240 mg/dL. Les corrélations d'échantillon non pondérées entre ces variables et la variable AGE sont 0,481 (PAS), 0,118 (PAD), 0,552 (HYP) et 0,060 (CTE), respectivement. Donc, il est raisonnable de postuler que la non-réponse totale en fonction de l'âge aura vraisemblablement des effets différents sur les estimateurs par sondage de ces quatre variables.

Au total, 4 727 personnes de l'ensemble de données de la NHANES étaient admissibles. Nous avons produit une non-réponse totale pour les quatre variables d'intérêt conformément à deux fonctions logistiques de la propension à répondre de la variable auxiliaire x prises comme étant l'âge (en années) du participant à l'enquête moins 18. Nous considérons dans ces fonctions les prédicteurs linéaire et

directe est que l'estimateur $\bar{y}^{\pi\psi v}$ n'est pas asymptotique-ment équivalent à $\bar{y}^{\pi\psi v}$ dans (2). Le corollaire qui suit fournit une distribution asymptotique pour $\bar{y}^{\pi\psi v}$, sous l'hypothèse de la normalité asymptotique de $\bar{y}^{\pi\psi v}$.

Corollaire 1. Supposons que les conditions du théorème 1 sont vérifiées. Supposons que le plan d'échantillonnage et le modèle de réponse sont tels que

$$\bar{y}^{\pi\psi v} - \bar{y}^{N_v} - B_v \frac{[\text{Var}(\bar{y}^{\pi\psi v})]^{1/2}}{\bar{y}^{\pi\psi v}} \xrightarrow{C} \mathcal{N}(0, 1) \text{ quand } v \rightarrow \infty,$$

où B_v est défini dans (13). Si, en outre,

$$\lim_{v \rightarrow \infty} (n_v h_v) \text{Var}(\bar{y}^{\pi\psi v}) \in (0, \infty),$$

alors

$$\frac{\bar{y}^{\pi\psi v} - \bar{y}^{N_v} - B_v \frac{[\text{Var}(\bar{y}^{\pi\psi v})]^{1/2}}{\bar{y}^{\pi\psi v}}}{\bar{y}^{\pi\psi v}} \xrightarrow{C} \mathcal{N}(0, 1).$$

Nous discutons maintenant des propriétés de la version

fondée sur le ratio de l'estimateur pondéré par la propension à réponse donné en (5). En nous basant sur les résultats obtenus pour $\bar{y}^{\pi\psi v}$, nous pouvons appliquer la théorie classique de l'estimation par le ratio pour calculer les résultats asymptotiques pour $\bar{y}^{\text{rat}, \pi\psi v}$. En particulier, sous les mêmes hypothèses, les taux asymptotiques pour le biais et la variance approximatifs de $\bar{y}^{\text{rat}, \pi\psi v}$ sont les mêmes que ceux indiqués au théorème 1, et la distribution asymptotique de $\bar{y}^{\text{rat}, \pi\psi v}$ est donnée dans le résultat qui suit.

Théorème 2. Supposons que les conditions du théorème 1 sont vérifiées. Supposons que la moyenne de population doit être estimée au moyen de l'estimateur pondéré par la propension à répondre $\bar{y}^{\text{rat}, \pi\psi v}$ donné en (5) et que les propensions à répondre sont estimées par $\hat{\phi}_i$, l'estimateur de régression par polynômes locaux de degré k défini en (8).

Soit

$$\bar{e}^{\pi\psi v} = \frac{1}{I} \sum_{i \in s_v} \pi_i^{-1} \hat{\psi}_i^{-1} (y_i - \bar{y}^{N_v}) R_i,$$

où les poids $\hat{\psi}_i^{-1}$ sont données au théorème 1. Supposons que

$$\frac{\bar{e}^{\pi\psi v} - E(\bar{e}^{\pi\psi v})}{\bar{e}^{\pi\psi v}} \xrightarrow{C} \mathcal{N}(0, 1) \text{ quand } v \rightarrow \infty,$$

et

$$\lim_{v \rightarrow \infty} (n_v h_v) \text{Var}(\bar{e}^{\pi\psi v}) \in (0, \infty).$$

Alors,

$$\frac{\bar{y}^{\text{rat}, \pi\psi v} - \bar{y}^{N_v} - B_{\text{rat}, v} \frac{[\text{Var}(\bar{e}^{\pi\psi v})]^{1/2}}{\bar{y}^{\text{rat}, \pi\psi v}}}{\bar{y}^{\text{rat}, \pi\psi v}} \xrightarrow{C} \mathcal{N}(0, 1)$$

quand $v \rightarrow \infty$, où $B_{\text{rat}, v} = O(h_v^{k+1/2})$, si k est impair, et $B_{\text{rat}, v} = O(h_v^{k+(3/2)})$, si k est pair.

4. Estimation de la variance

Comme nous l'avons mentionné à la section 3, l'estimateur $\bar{y}^{\pi\psi v}$ n'est pas asymptotiquement équivalent à $\bar{y}^{\pi\psi v}$ de sorte qu'il est habituellement incorrect d'approximer la variance de $\bar{y}^{\pi\psi v}$ est donnée par Kim et Kim (2007) quand on émet l'hypothèse que les propensions à répondre suivent un modèle paramétrique. Dans le présent contexte, la variance asymptotique de $\bar{y}^{\pi\psi v}$ est

$$\text{Var}[\bar{y}^{\pi\psi v}] = \text{Var}\left[\frac{1}{I} \sum_{i \in s_v} \pi_i^{-1} \hat{\psi}_i^{-1} R_i y_i\right],$$

avec $\hat{\psi}_i^{-1}$ donné par le théorème 1. Nous avons souligné antérieurement dans Da Silva et Opsomer (2006), pour le cas plus simple d'un polynôme de degré nul, que la complexité élevée de l'expression rend l'estimation directe de cette variance difficilement applicable et avons proposé une méthode de rééchantillonnage pour la remplacer. Nous

décrivons brièvement la procédure ici en l'étendant aux polynômes locaux de degré k , mais nous omettons les calculs théoriques.

Nous partons d'un ensemble de poids de rééchantillonnage en l'absence de non-réponse, défini par l'estimation de la variance d'un estimateur linéaire.

$$\hat{\theta} = \frac{1}{I} \sum_{i \in s_v} w_i y_i,$$

L'estimateur de variance par rééchantillonnage de $\hat{\theta}$ est défini comme étant

$$\hat{V}(\hat{\theta}) = \sum_{\ell=1}^L c_{\ell}^2 (\hat{\theta}_{(\ell)} - \hat{\theta})^2, \quad (17)$$

où

$$\hat{\theta}_{(\ell)} = \frac{1}{I} \sum_{i \in s_v} w_i^{(\ell)} y_i, \quad \ell = 1, 2, \dots, L_v,$$

désigne un ensemble de L_v répliques pour $\hat{\theta}$, $w_i^{(\ell)}$ représente les poids d'échantillonnage associés à la ℓ^{e} réplique et c_{ℓ} est un facteur qui dépend de la méthode de rééchantillonnage. Les exemples de procédures de rééchantillonnage satisfaisant l'équation (17) s'appuient sur des variantes de la méthode du jackknife ou de la technique des répétées équilibrées. Le processus d'adaptation de la méthode de rééchantillonnage à l'estimation de la variance de $\bar{y}^{\pi\psi v}$ et de $\bar{y}^{\text{rat}, \pi\psi v}$ est simple. Les répliques nécessaires de ces estimateurs corrigés, à savoir $\bar{y}^{(\ell)}$ et $\bar{y}^{\text{rat}, \pi\psi v, (\ell)}$, sont obtenues en remplaçant les $w_i = \pi_i^{-1}$ par $w_i^{(\ell)}$ dans (4) et (5), respectivement, ainsi que dans les calculs requis pour produire la fonction $\hat{\phi}_i$ dans (9). À la section 5.4 qui suit, nous évaluons les propriétés pratiques de la méthode d'estimation de la variance par rééchantillonnage en appliquant aux données de la NHANES.

lequel la population U_v est emboîtée dans la série croissante de populations $\{U_v : N_v < N_{v+1}^{\infty} = \infty\}$. À partir de chaque population U_v , nous tirons un échantillon s_v de taille $n_v(n_v \geq n_{v-1})$ selon un plan d'échantillonnage $P_v(\cdot)$. Ce cadre est habituellement adopté dans les études asymptotiques des estimateurs par sondage. Voir Isaki et Fuller (1982) pour l'une des premières références.

À titre d'approximation de $\phi_i = \phi(x_i)$ fondée sur la population, nous considérerons dans le calcul de la plupart des résultats présentés ici l'ajustement en population par la régression par polynômes locaux.

$$\hat{\phi}_i \equiv \hat{\phi}(x_i, k, h_v) = e_i' \mathbf{B}_i = e_i' \mathbf{T}_i^{-1} \mathbf{t}_i, \quad i \in U_v. \quad (10)$$

où

$$(\mathbf{T}_i, \mathbf{t}_i) \equiv (\{\mathbf{T}_i^{(p)}\}_{k+1}^{pd}, \{\mathbf{t}_i^{(p)}\}_{k+1}^{pd=1})$$

$$\equiv E(\mathbf{T}_i^{(s)}, \mathbf{t}_i^{(s)}) = (\mathbf{X}_i' \mathbf{W}_i \mathbf{X}_{-i}^{(s)}, \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_{-i}^{(s)} \mathbf{W}_i^{(U)} \boldsymbol{\Phi}^{(U)})$$

Les résultats (12) et (13) impliquent que l'estimateur pondéré par la propension à répondre $\bar{y}^{\pi_{\psi v}}$, en utilisant un estimateur de la propension à répondre basé sur la régression par polynômes locaux, est asymptotiquement sans biais pour la moyenne de population \bar{y}^{N_v} sous la distribution conjointe du plan d'échantillonnage et du modèle de réponse (1). En combinant ce résultat avec (14), nous obtenons

$$\bar{y}^{\pi_{\psi v}} = \bar{y}^{N_v} + O_p \left(\frac{\sqrt{n_v h_v}}{1} \right), \quad \text{quand la largeur de fenêtre satisfait} \quad (15)$$

$$h_v = \begin{cases} O \left(n_v^{\frac{1}{2k+4}} \right), & k \text{ pair,} \\ O \left(n_v^{\frac{1}{2k+3}} \right), & k \text{ impair.} \end{cases} \quad (16)$$

Donc, sans émettre l'hypothèse d'une forme paramétrique pour la fonction de propension à répondre $\phi(\cdot)$, $\bar{y}^{\pi_{\psi v}}$ est convergent pour la moyenne de population sous le plan d'échantillonnage et sous le modèle de réponse, à condition que les propensions à répondre soient une fonction lisse de la covariable x . Le prix de cette robustesse est que le taux de convergence est d'ordre $\sqrt{n_v h_v}$ au lieu du taux paramétrique habituel $\sqrt{n_v}$. Cependant, à mesure qu'augmente le degré k des polynômes locaux, le taux de convergence s'améliore. Puisque l'estimateur de régression par la méthode du noyau donne dans Da Silva et Opsomer (2006) équivalent au cas $k = 0$, la régression par polynômes locaux de degré plus élevé est asymptotiquement supérieure à la régression par la méthode du noyau dans le contexte d'un ajustement pour tenir compte de la non-réponse. Cette constatation théorique concorde avec celle faite dans d'autres contextes (voir, par exemple, Wand et Jones 1995, page 130).

L'expression (11) du théorème 1 généralise une autre constatation faite par Da Silva et Opsomer (2006) au cas de la régression par polynômes locaux, à savoir que les poids asymptotiques ψ_i^{-1} ne peuvent pas être approximés par l'inverse des propensions à répondre ϕ_i^{-1} (ou leurs estimations $\hat{\phi}_i^{-1}$). Une conséquence

$$\bar{y}^{\pi_{\psi v}} = \frac{1}{N} \sum_{i \in s_v} \pi_i^{-1} \psi_i^{-1} y_i R_i, \quad (11)$$

où

$$\psi_i^{-1} = \phi_i^{-1} - \phi_i^{-2} e_i' \mathbf{T}_i^{-1} (\mathbf{t}_i^{(s)} - \mathbf{T}_i^{(s)} \mathbf{B}_i),$$

$\mathbf{t}_i^{(s)}$ et $\mathbf{T}_i^{(s)}$ sont donnés dans (7) et $\hat{\phi}_i, \mathbf{B}_i, \mathbf{T}_i$ sont définis dans (10). Alors,

$$E[(\bar{y}^{\pi_{\psi v}} - \bar{y}^{\pi_{\psi v}})^2] = O \left(\frac{n_v^2 h_v^2}{1} \right) \quad \text{et le biais et la variance de } \bar{y}^{\pi_{\psi v}} \text{ satisfont} \quad (12)$$

Afin d'appliquer les estimateurs pondérés par la propension à répondre (4) et (5), il faut estimer les propensions ϕ_i . Dans Da Silva et Opsomer (2006), nous avons recouru à la régression par la méthode du noyau, qui peut être décrite comme il suit. Soit $K(\cdot)$ une fonction continue et positive et h_y sa largeur de fenêtre. Définissons la matrice de dimensions $N_y \times (k + 1)$

$$\mathbf{X}^{U_i} = \begin{bmatrix} 1 & (x_1 - x_i) & \dots & (x_1 - x_i)^k \\ \vdots & \vdots & & \vdots \\ 1 & (x_{N_y} - x_i) & \dots & (x_{N_y} - x_i)^k \end{bmatrix},$$

la matrice de dimensions $N_y \times N_y$

$$\mathbf{W}^{U_i} = \text{diag} \left\{ \frac{1}{K} \frac{h_y}{x_j - x_i} : 1 \leq j \leq N_y \right\},$$

et le vecteur de population des indicateurs de réponse $\mathbf{R}^{U_i} = (R_1^i, R_2^i, \dots, R_{N_y}^i)$. Le vecteur \mathbf{R}^{U_i} serait connu si, au lieu de l'échantillon s_y , nous envisagions un recensement de la population U_y . Dans un tel cas, l'estimateur par la régression par polynômes locaux de degré k de $\phi_i = \phi(x_i)$, basé sur l'ensemble de la population, serait donné par le modèle

$$\phi_i^{U_i} = e_i^1 (\mathbf{X}^{U_i} \mathbf{W}^{U_i} \mathbf{X}^{U_i})^{-1} \mathbf{X}^{U_i} \mathbf{W}^{U_i} \mathbf{R}^{U_i} \quad (6)$$

où e_j désigne la j^{e} colonne de la matrice identité d'ordre $k + 1$ et où $\mathbf{X}^{U_i} \mathbf{W}^{U_i} \mathbf{X}^{U_i}$ est supposée non singulière.

Puisque les valeurs des indicateurs de réponse ne sont observées que pour les unités sélectionnées dans l'échantillon, l'ajustement du modèle à la population donné en (6) est irréalisable. Cependant, si nous définissons \mathbf{X}^{s_y} comme étant la matrice de dimensions $n_y \times (k + 1)$ formée par les lignes de \mathbf{X}^{U_i} correspondant aux unités $j \in s_y$,

$$\mathbf{W}^{s_y} = \text{diag} \left\{ \frac{1}{K} \frac{h_y}{x_j - x_i} : j \in s_y \right\}$$

et $\mathbf{R}^{s_y} = (R_j : j \in s_y)$, un estimateur par la régression par polynômes locaux de degré k basé sur l'échantillon de $\phi_i = \phi(x_i)$ est donné par

$$\phi_i^0 = e_i^1 \hat{\mathbf{T}}_{-1}^{-1} \hat{\mathbf{T}}^{s_y} \quad (7)$$

où

$$(\hat{\mathbf{T}}^{s_y}, \hat{\mathbf{T}}^{s_y}) \equiv \left(\hat{\mathbf{T}}^{s_y, d=1, \dots, k+1}, \hat{\mathbf{T}}^{s_y, d=1, \dots, k+1} \right) = (\mathbf{X}^{s_y} \mathbf{W}^{s_y} \mathbf{X}^{s_y}, \mathbf{X}^{s_y} \mathbf{W}^{s_y} \mathbf{R}^{s_y})$$

et $\hat{\mathbf{T}}^{s_y}$ est supposée inversible. Nous obtenons un cas particulier de (7) en considérant $k = 0$, qui correspond à

$$\hat{\phi}(x_i, k, h_y) = e_i^1 \left(\hat{\mathbf{T}}^{s_y} + \text{diag} \left\{ \frac{\delta_1}{N_y} \right\} \right)^{-1} \hat{\mathbf{T}}^{s_y} \quad (8)$$

où δ_1 est une petite constante positive. Les termes d'ordre faible δ_1/N_y ajoutés à la diagonale principale de $\hat{\mathbf{T}}^{s_y}$ suffisent pour rendre la matrice ajustée résultante inversible pour tout h_y . Par conséquent, $\hat{\phi}(x_i, k, h_y)$ sera bien défini, pour tout $i \in s_y$. Cependant, l'utilisation de $\hat{\phi}(x_i, k, h_y)$ comme correction de la pondération par la propension à répondre pose une autre difficulté technique, parce que l'estimateur de la propension à répondre (8) peut effectivement devenir arbitrairement proche de zéro. Pour résoudre ce problème, nous rendons $\hat{\phi}(x_i, k, h_y)$ strictement non nul en considérant l'estimateur

$$\hat{\phi}_i = \max \{ \hat{\phi}(x_i, k, h_y), \delta_2 (N_y h_y)^{-1} \}, \quad (9)$$

pour une certaine constante $\delta_2 > 0$. Cette idée est reliée à l'estimateur de régression par la méthode du noyau.

3. Propriétés asymptotiques

À la présente section, nous présentons les propriétés des estimateurs pondérés par la propension à répondre (4) et (5) quand les propensions sont estimées par la régression par polynômes locaux (9). Les hypothèses, lemmes et exposés des preuves pour les résultats qui suivent figurent en annexe, et une étude théorique complète peut être consultée dans Da Silva et Opsomer (2008). Nous ne présentons pas les calculs complets ici, parce qu'ils suivent l'approche générale décrite dans Da Silva et Opsomer (2006). Nous considérons un cadre asymptotique suivant

l'estimateur de la régression par la méthode du noyau décrit dans Da Silva et Opsomer (2006). D'autres cas particuliers de (7) sont les estimateurs linéaire local, quadratique local et cubique local de la propension à répondre, qui résultent de l'ajustement local de polynômes de premier, deuxième et troisième degré, respectivement.

En pratique, quand $\hat{\mathbf{T}}^{s_y}$ est singulière, une procédure simple pour s'assurer que ϕ_i^0 soit bien défini consiste à choisir une largeur de fenêtre suffisamment grande pour garantir qu'au moins $k + 1$ valeurs de R_j soient comprises dans la fenêtre $[x_i - h_y, x_i + h_y]$, pour tout $i \in s_y$. Si cette fenêtre ne contient pas suffisamment d'indicateurs de réponse et que la largeur de fenêtre doit demeurer fixe, il convient d'envisager une autre approche. À cette fin, nous adoptons ici l'ajustement fait par Breidt et Opsomer (2000) et définissons l'estimateur par la régression par polynômes locaux de degré k fondé sur l'échantillon de $\phi_i = \phi(x_i)$ par

où la forme exacte de la *fonction de propension à répondre* $\phi(\cdot)$ n'est pas spécifiée, mais on suppose qu'il s'agit d'une fonction lisse de x_i avec $\phi(\cdot) \in (0, 1]$. La relation (1) définit un processus de non-réponse dit *indépendant*, en ce sens que les propensions à répondre sont indépendantes des valeurs de toute variable étudiée, sachant la covariable x (voir Lohr 1999, page 265). Par conséquent, la théorie élaborée ici n'est pas destinée à traiter les mécanismes de réponses non ignorables.

Si toutes les propensions à répondre étaient connues, les corrections de pondération résultantes pourraient être obtenues en appliquant une approche d'estimation en deux phases. Par exemple, deux estimateurs possibles de la moyenne de population \bar{y}_{N_v} seraient donnés par

$$(2) \quad \bar{y}^{\pi\phi v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} y_i R_i$$

et

$$(3) \quad \bar{y}^{\pi t, \pi\phi v} = \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} y_i R_i / \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} R_i$$

qui sont des formes d'ajustement appliquées aux estimateurs d'Horvitz-Thompson et de Hájek en vue de tenir compte de la non-réponse totale. La même idée peut être utilisée pour obtenir des ajustements par pondération par la propension à répondre applicables à l'estimateur par la régression généralisée pour l'estimation en présence de non-réponse (Cassel et coll. 1983).

Les estimateurs (2) et (3) sont sans biais et presque sans biais pour \bar{y}_{N_v} respectivement, sous l'approche de quasi randomisation de Oh et Schuren (1983), où les propriétés statistiques sont évaluées en utilisant la distribution conjointe sous le plan d'échantillonnage et sous le modèle de réponse. Cependant, en pratique, les propensions à répondre sont habituellement inconnues, et, dans (2) et (3), nous devons remplacer la fonction ϕ_i par des estimations $\hat{\phi}_i$, qui satisfont $0 < \hat{\phi}_i \leq 1$. Les estimateurs résultants pondérés par la propension à répondre sont par conséquent

$$(4) \quad \bar{y}^{\pi\hat{\phi} v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} y_i R_i$$

et

$$(5) \quad \bar{y}^{\pi t, \pi\hat{\phi} v} = \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} y_i R_i / \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} R_i$$

La dernière formule a l'avantage d'être invariante au changement de position et d'échelle, parce que la somme de ses poids corrigés $\pi_i^{-1} \hat{\phi}_i^{-1} R_i / \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} R_i$ est égale à un et qu'elle ne requiert pas que la taille de population N_v soit connue.

locaux, une technique non paramétrique décrite, par exemple, dans Wand et Jones (1995). Comparativement au lissage par la méthode du noyau, la régression par polynômes locaux améliore l'approximation locale de la fonction de propension à répondre inconnue, ce qui donne de meilleures propriétés pratiques et théoriques. Elle est également utilisée beaucoup plus fréquemment comme méthode de lissage en pratique, et est implémentée dans la plupart des grands programmes statistiques. Deuxièmement, nous appliquons la méthode non paramétrique d'estimation des scores de propension aux données de la National Health and Nutrition Examination Survey (NHANES), ce qui nous permet de comparer plusieurs méthodes de correction de la non-réponse, paramétriques ainsi que non paramétriques, dans des conditions réalistes.

À la section 2, nous présentons la méthode de pondération et l'estimation des propensions à répondre. À la section 3, nous discutons des propriétés théoriques des estimateurs corrigés. À la section 4, nous décrivons comment adapter une méthode d'estimation de la variance par ré-échantillonnage pour estimer la variance des estimateurs corrigés. Enfin, à la section 5, nous démontrons les propriétés en échantillon fini des estimateurs au moyen d'une expérience par simulation portant sur des données provenant de la NHANES.

2. Pondération par la régression par polynômes locaux

Considérons une population de N_v unités, désignée par $U_v = \{1, 2, \dots, N_v\}$. Supposons qu'un échantillon s_v est tiré de U_v selon un plan d'échantillonnage probabiliste $p(s_v)$. Soit n_v la taille de s_v et $\pi_i = \Pr\{i \in s_v\} = \sum_{i \in s_v} p(s_v)$ la probabilité d'inclusion de l'unité i , pour tout $i \in U_v$. Nous souhaitons estimer la moyenne de population d'une variable étudiée y_i à savoir $\bar{y}_{N_v} = N_v^{-1} \sum_{i \in U_v} y_i$, où y_i désigne la valeur de y pour la i^{e} unité de U_v . Nous supposons que les valeurs x_i d'une variable auxiliaire x sont entièrement observées dans l'échantillon. Soit $y^v = (y_1^v, \dots, y_{N_v}^v)$ et l'expression comparable pour x^v .

Si l'existe des cas de non-réponse totale dans l'échantillon, nous n'observons les valeurs des variables étudiées que pour les unités comprises dans un sous-ensemble $r_v \subset s_v$. Afin de tenir compte de l'information perdue pour estimer les paramètres d'intérêt, il devient nécessaire de modéliser le processus de réponse. Pour définir ce modèle de réponse, posons que R_i est une variable indicatrice qui prend la valeur un si l'unité i répond à l'enquête et la valeur zéro autrement, pour tout $i \in s_v$. Nous supposons que, étant donné l'échantillon, les indicateurs de réponses sont des variables aléatoires bernoulliennes indépendantes avec

$$(1) \quad \Pr\{R_i = 1 | i \in s_v, y^v, x^v\} = \phi(x_i) \equiv \phi_i, \text{ pour tout } i \in s_v,$$

Pondération par la propension à répondre non paramétrique fondée sur la régression par polynômes locaux pour corriger la non-réponse aux enquêtes

Damião N. da Silva et Jean D. Opsomer¹

Résumé

La pondération par la propension à répondre est une méthode de rajustement pour tenir compte de la non-réponse totale dans les enquêtes. Une forme de mise en œuvre de cette méthode consiste à diviser les poids d'échantillonnage par les estimations de la probabilité que les unités échantillonnées répondent à l'enquête. Habituellement, ces estimations sont obtenues par rajustement de modèles paramétriques, tels qu'une régression logistique. Les estimateurs corrigés résultants peuvent devenir biaisés si les modèles paramétriques sont spécifiés incorrectement. Afin d'éviter les erreurs de spécification du modèle, nous considérons l'estimation non paramétrique des probabilités de réponse par la régression par polynômes locaux. Nous étudions les propriétés asymptotiques de l'estimateur résultant sous quasi randomisation. Nous évaluons en pratique le comportement de la méthode proposée de correction de la non-réponse en nous servant de données de la NHANES.

Mots clés : Régression par la méthode du noyau ; données manquantes ; scores de propension ; non-réponse totale ; pondération.

1. Introduction

La pondération par la propension à répondre est une méthode souvent appliquée dans les enquêtes par sondage pour tenir compte de la non-réponse totale. Sous ce type de non-réponse, la collecte de données complètes n'est effectuée qu'après d'une partie des unités sélectionnées dans l'échantillon, auxquelles on donne le nom de répondants. La méthode de pondération par la propension à répondre consiste à accroître les poids d'échantillonnage des répondants présents dans l'échantillon en se servant d'estimations de la probabilité qu'ils répondent à l'enquête. Cette probabilité est également appelée propension à répondre en raison de son analogie avec la théorie des scores de propension élaborée par Rosenbaum et Rubin (1983) pour les études par observation et intégrée dans les problèmes de non-réponse aux enquêtes par David, Little, Samuël et Triest (1983). Des descriptions générales de la pondération par la propension à répondre pour corriger de la non-réponse les estimateurs par sondage classiques peuvent être consultées, par exemple, dans Nargundkar et Joshi (1975), Cassel, Samdal et Wretman (1983), ainsi que Groves, Dillman, Eltinge et Little (2002). Dans la mise en œuvre habituelle de la méthode, les probabilités de réponses sont estimées au moyen de courbes de régression paramétriques, comme des modèles logistiques, probit ou exponentiels. Voir à cet égard Alho (1990), Folsom (1991), Ekholm et Laaksonen (1991) ainsi que l'annexe, Milne et Folsom (1991). Un compte rendu théorique récent des propriétés statistiques de la méthode est donné dans Kim et Kim (2007). Ces modèles paramétriques sont ajustés facilement sous forme de modèles linéaires généralisés. Cependant, un aspect important, et parfois oublié, de cette méthode est la spécification de la forme de la fonction de lien qui relie les propensions à répondre à un prédicteur linéaire de l'information auxiliaire. Si cette fonction, que nous appellerons fonction de propension à répondre, est spécifiée incorrectement, les estimateurs corrigés résultants des quantités de population seront probablement biaisés. Les méthodes non paramétriques offrent une autre approche d'estimation des propensions à répondre. La forme paramétrique de la fonction de propension à répondre n'a pas à être spécifiée. Ces méthodes offrent donc, en ce sens, une solution attrayante quand on veut contourner le choix d'une fonction de lien, comme l'a souligné Laaksonen (2006), ou quand un modèle paramétrique est difficile à spécifier a priori. Dans ce contexte, Gionmi (1984) a proposé d'utiliser le lissage par la méthode du noyau, sous la forme de l'estimateur de Nadaraya-Watson, pour estimer les probabilités de réponse. Da Silva et Opsomer (2006) ont établi la convergence de l'estimateur de Gionmi pour la moyenne de population ainsi que les taux de dérivées pour le biais asymptotique et la variance. Les propriétés théoriques d'un estimateur de variance par la méthode du jackknife ont également été étudiées. Dans le présent article, nous étendons dans deux directions les résultats présentés dans Da Silva et Opsomer (2006). Premièrement, nous considérons l'estimation des propensions à répondre par la régression par polynômes

1. Damião N. da Silva, Departamento de Estatística, Campus Universitário, Natal, RN 59078-970, États-Unis. Courriel : jopsomer@stat.colostate.edu; Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, États-Unis. Courriel : damiao@ccet.uflm.br;

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquêtes*, 21, 27-35.

Lavallée, P. (2001). Correcting for non-response in indirect sampling. *Proceedings of Statistics Canada's Symposium 2001*.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles et Éditions Ellipse.

Lavallée, P. (2007). *Indirect Sampling*. New York : Springer.

Lévesque, I. (2001). Enquête sur la dynamique du travail et du revenu - Estimation de la variance. Rapport interne de Statistique Canada, 2 juillet 2001.

Sämdal, C.-E., Swensson, B. et Wrethman, J. (1991). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Tableau 10
Comparaison des estimations de l'écart-type

Variables	Revenu total avant	Revenu total après	Gains	Salaires et traitements avant	Salaires et traitements après
Niveau national	9 677 258 789	7 343 792 762	8 850 202 075	8 468 718 449	8 232 428 642
Ontario	Méthode (2) sans liens manquants	7 888 106 377	6 101 001 739	7 245 688 373	7 149 203 530
	Méthode (2) avec moyenne	7 601 169 501	5 939 509 894	6 952 217 872	6 831 300 511
Québec	MCGP sans liens manquants	4 341 215 711	3 113 247 130	3 772 369 180	3 162 277 660
	Méthode (2) avec moyenne	4 160 251 472	2 974 248 451	3 668 996 929	3 100 868 366

5. Conclusion

Nous avons élaboré quatre méthodes d'estimation pour traiter le problème de la non-réponse de lien dans l'échantillonnage indirect. Les résultats des simulations exposés dans l'article montrent que les méthodes de correction que nous avons présentées pour illustrer l'utilisation de la MGP avec intégration de la non-réponse de lien donnent de bons résultats pour ce qui est de réduire le biais de l'estimation et produisent une amélioration globale de la variance. Le progrès en ce qui concerne la réduction du biais semble important. L'application de la méthode proposée à la section 3.2 à des ensembles de données réelles sera étudiée prochainement.

Les constatations importantes qui se dégagent de la présente étude sont les suivantes :

1. Les méthodes d'ajustement sont faciles à appliquer.
2. Dans une situation plus générale, telle que $L_{ji} > 1$ pour certaines unités j , (35) représente la moyenne pondérée par L_{ji}^* . Conséquemment, l'approche de la médiane donnée par (39) et (40) peut être modifiée en utilisant une version généralisée de la médiane, c'est-à-dire la médiane « pondérée ».

Autrement dit, nous remplaçons (38) par

$$w_{j,\text{médian}} = \frac{1}{\sum_j \pi_j}$$

$$\text{ou } j = 1, 2, \dots, L_B^*, 1, 2, \dots, L_B^*, \dots, 1, 2, \dots, L_B^*, \dots, 1, 2, \dots, L_B^*, \dots$$

3. Certaines réponses de lien valides en dehors de s^* ne peuvent pas être utilisées pour estimer L_B^* par les méthodes proposées à la section 3.1. Toutefois, cette information valide pourrait être avantageuse dans les approches de prédiction de L_{jik} en utilisant des variables auxiliaires, comme le montre la section 3.2.1.

Cassey, C.-M., Samdal, C.-E. et Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York : John Wiley & Sons, Inc.

Deville, J.-C., et Lavalée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 185-196.

Draeger, N.R., et Smith, H. (1998). *Applied Regression Analysis*, 3^{ème} Ed. New York : John Wiley & Sons, Inc.

Ernst, L. (1989). Weighing issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York : John Wiley & Sons, Inc., 135-159.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Hurand, C. (2006). La méthode généralisée du partage des poids et le problème d'identification des liens. Rapport interne, Division des méthodes d'enquêtes sociales, Statistique Canada, juillet 2006.

Kalton, G., et Brack, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-49.

LaRoche, S. (2003). Longitudinal and Cross-Sectional Weighing of the Survey of Labour and Income Dynamics. *Income Research Paper Series*, Catalogue no. 75F0002MIE - No. 007, Statistique Canada.

Lavalée, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. *Statistics Canada, Research Paper of the Survey of Labour and Income Dynamics*, Catalogue No. 93-19, décembre 1993.

Bibliographie

Les auteurs remercient le rédacteur associé et deux examinateurs de leurs suggestions et commentaires constructifs au sujet des versions précédentes du présent article. Ces travaux de recherche ont été financés par le Conseil de recherches en sciences naturelles et en génie du Canada et par Mathématiques des technologies de l'information et des systèmes complexes.

Remerciements

L'EDTR. Le gain en ce qui concerne la réduction de la variance n'est pas aussi important que dans le cas du biais; toutefois, l'étude par simulation montre que la méthode proposée produit une plus petite variance que l'application de la MGPF sans correction pour les liens manquants. Les résultats des simulations présentés ici sont fondés sur un seul échantillon de l'EDTR et une seule suppression aléatoire des liens des individus initialement présents. Pour évaluer complètement les propriétés des estimateurs susmentionnés, nous aurions dû utiliser un processus Monte Carlo. Des simulations de ce type ont été effectuées par Hurand (2006) en se basant sur des données agricoles. Dans ces simulations, 1 000 échantillons ont été sélectionnés et, pour chaque échantillon sélectionné, le pire scénario a été utilisé, c'est-à-dire l'élimination de tous les liens provenant des unités non échantillonnées. Les résultats de ces simulations ont montré que l'ajustement proportionnel et l'ajustement proportionnel global sont les deux méthodes qui s'approchent, en moyenne, le plus du total réel et celles dont les biais sont négligeables.

Tableau 8
Comparaison des erreurs relatives dans les estimations des gains (%)

Province	MGPF avec liens manquants	Méthode (1) avec moyenne	Méthode (1) avec médiane	Méthode (2) avec moyenne	Méthode (2) avec médiane
T.-N.-L.	6,286	1,682	1,509	0,041	3,585
I.-P.-E.	6,382	3,470	3,397	0,0739	7,115
N.-E.	8,115	1,773	1,308	1,265	5,281
N.-B.	6,930	1,430	1,062	1,279	4,512
Qc	8,472	0,773	0,706	2,827	4,560
Ont.	6,955	0,336	0,213	3,760	2,920
Man.	8,172	2,879	2,512	0,291	5,835
Sask.	7,793	1,467	1,055	0,979	4,324
Alb.	8,739	0,691	0,475	2,140	3,777
C.-B.	6,553	1,091	0,914	2,643	5,081
Canada	7,522	0,715	0,664	2,628	4,131

Tableau 9
Comparaison des erreurs relatives dans les estimations des traitements et salaires avant retenues (%)

Province	MGPF avec liens manquants	Méthode (1) avec moyenne	Méthode (1) avec médiane	Méthode (2) avec moyenne	Méthode (2) avec médiane
T.-N.-L.	6,336	1,656	1,484	0,1012	3,593
I.-P.-E.	6,809	4,734	4,694	1,056	8,424
N.-E.	8,231	2,047	1,573	0,939	5,509
N.-B.	7,131	2,036	1,664	0,685	5,133
Qc	8,704	1,278	1,162	2,294	5,070
Ont.	7,096	0,0317	0,0791	3,473	3,265
Man.	8,404	3,436	3,065	0,787	6,469
Sask.	8,202	2,353	1,953	0,107	5,213
Alb.	9,059	1,142	0,918	1,713	4,247
C.-B.	6,547	1,190	0,992	2,565	5,234
Canada	7,699	1,098	1,025	2,251	4,541

Tableau 5
Salaires et traitements avant retenues (en dollars canadiens)

Province	liens manquants		liens manquants	
	Estim. par MGPP sans	Estim. par MGPP avec	Estim. par MGPP sans	Estim. par MGPP avec
T.-N.-L.	6 180 713 343	6 572 345 010	1 747 755 878	1 713 809 312
I.-P.-E.	1 636 344 440	1 341 912 666	12 579 519 733	10 961 105 589
N.-B.	10 742 381 379	11 508 445 078	1,10024E+11	2,07265E+11
Qc	1,08636E+11	1,18092E+11	16 701 823 718	16 411 467 435
Ont.	2,07331E+11	2,22043E+11	53 195 227 508	56 875 663 895
Man.	16 146 993 217	17 504 024 442	56 764 297 512	4,90763E+11
Sask.	13 982 423 360	15 129 217 320	53 077 388 907	
Alb.	52 594 490 290	57 359 188 114		
C.-B.	56 206 787 033	59 886 429 369		
Canada	4,85784E+11	5,23184E+11		

Tableau 6
Comparaison des erreurs relatives dans l'estimation du revenu avant impôt (%)

Province	MGPP avec liens manquants		Méthode (1) avec moyenne	
	Méthode (1) avec	Méthode (1) avec	Méthode (1) avec	Méthode (2) avec
T.-N.-L.	5,688	0,599	0,460	1,059
I.-P.-E.	5,075	0,570	0,532	2,859
N.-B.	6,433	0,037	0,397	2,693
Qc	7,492	0,828	0,909	4,372
Ont.	6,267	1,413	1,348	4,691
Man.	6,836	0,838	0,500	1,444
Sask.	6,988	0,107	0,446	2,480
Alb.	7,982	0,502	0,696	3,185
C.-B.	5,926	0,442	0,612	3,995
Canada	6,734	0,650	0,724	3,868

Tableau 7
Comparaison des erreurs relatives dans l'estimation du revenu après impôt (%)

Province	MGPP avec liens manquants		Méthode (1) avec moyenne	
	Méthode (1) avec	Méthode (1) avec	Méthode (1) avec	Méthode (2) avec
T.-N.-L.	5,617	0,588	0,457	1,101
I.-P.-E.	5,061	0,616	0,585	2,832
N.-B.	6,522	0,127	0,226	2,539
Qc	7,742	0,403	0,496	3,991
Ont.	6,387	1,130	1,068	4,432
Man.	6,836	0,881	0,551	1,451
Sask.	6,977	0,027	0,356	2,406
Alb.	7,984	0,493	0,684	3,180
C.-B.	5,909	0,406	0,584	3,989
Canada	6,841	0,415	0,492	3,657

Tableau 2
Revenu total avant impôt (en dollars canadiens)

Province	Estim. par MGPP sans	Estim. par MGPP avec	Estim. par MGPP corrigée	Estim. par MGPP corrigée avec médiane
T.-N.-L.	9 261 958 108	9 788 749 735	9 317 420 236	9 304 530 248
I.-P.-É.	2 720 448 008	2 858 506 466	2 735 943 043	2 734 922 451
N.-É.	18 277 017 251	19 573 546 299	18 140 076 618	18 067 144 557
N.-B.	15 297 155 323	16 281 178 934	15 291 696 585	15 236 482 035
Qc	1,378399E+11	1,69664E+11	1,56533E+11	1,56405E+11
Ont.	2,895E+11	3,07642E+11	2,85409E+11	2,85599E+11
Man.	23 436 397 548	25 043 168 032	23 632 717 226	23 553 543 216
Sask.	20 185 285 649	21 595 804 296	20 163 683 598	20 095 359 071
Alb.	69 063 402 292	74 576 351 600	68 716 661 193	68 582 541 735
C.-B.	81 749 374 346	86 593 614 506	81 387 640 982	81 248 680 715
Canada	6,8733E+11	7,33617E+11	6,8286E+11	6,82356E+11

Tableau 3
Revenu total avant impôt (en dollars canadiens)

Province	Estim. par MGPP sans	Estim. par MGPP avec	Estim. par MGPP corrigée	Estim. par MGPP corrigée avec médiane
T.-N.-L.	7 846 587 557	8 287 351 908	7 892 754 014	7 882 437 105
I.-P.-É.	2 300 092 795	2 416 503 441	2 314 256 124	2 313 544 320
N.-É.	15 154 508 564	16 257 679 161	15 080 155 194	15 020 088 623
N.-B.	12 878 350 198	13 718 260 686	12 894 700 593	12 849 252 205
Qc	1,27632E+11	1,37514E+11	1,27118E+11	1,26999E+11
Ont.	2,3788E+11	2,53073E+11	2,35192E+11	2,3534E+11
Man.	19 541 510 220	20 877 377 918	19 713 628 649	19 649 142 217
Sask.	16 894 929 025	18 073 635 883	16 890 410 993	16 834 787 407
Alb.	57 466 974 767	62 055 315 246	57 183 814 491	57 073 904 623
C.-B.	68 710 569 670	72 770 595 462	68 431 531 373	68 309 055 749
Canada	5,66306E+11	6,05044E+11	5,63958E+11	5,63518E+11

Tableau 4
Gains (en dollars canadiens)

Province	Estim. par MGPP sans	Estim. par MGPP avec	Estim. par MGPP corrigée	Estim. par MGPP corrigée avec médiane
T.-N.-L.	6 433 112 169	6 837 522 157	6 541 306 193	6 530 174 122
I.-P.-É.	1 898 192 704	2 019 341 995	1 964 066 449	1 962 669 664
N.-É.	12 772 667 160	13 809 197 160	12 999 111 234	12 939 785 579
N.-B.	11 250 688 811	12 030 378 710	11 411 530 716	11 370 222 533
Qc	1,18878E+11	1,28949E+11	1,19797E+11	1,19717E+11
Ont.	2,27577E+11	2,43404E+11	2,26812E+11	2,27092E+11
Man.	17 560 695 670	18 995 682 322	18 066 353 153	18 001 882 362
Sask.	15 159 319 031	16 340 668 148	15 381 733 004	15 319 210 228
Alb.	56 152 023 359	61 059 244 608	56 540 145 524	56 418 889 147
C.-B.	60 532 655 979	64 499 398 960	61 192 920 832	61 085 986 951
Canada	5,28214E+11	5,67945E+11	5,3199E+11	5,31722E+11

(35)

$$w_{i,\text{moyen}}^j = \frac{\sum_{j=1}^m L_{ji,i} \pi_j^A}{\sum_{j=1}^m L_{ji,i}} = \frac{\sum_{j=1}^m L_{ji,i}}{\sum_{j=1}^m \pi_j^A}$$

Nous écrivons

(36)

$$\hat{Y}_{B(1)}^j = \sum_{i=1}^n m_i^A \frac{L_{ji,i}}{w_{i,\text{moyen}}^j} \sum_{k=1}^K Y_{ik}^j$$

(37)

$$\hat{Y}_{B(2)}^j = m_i^A \sum_{i=1}^n w_i^j \sum_{k=1}^K Y_{ik}^j$$

respectivement. Notons que $w_{i,\text{moyen}}^j$ est le poids moyen des personnes longitudinales qui vivent dans le i^{e} ménage durant l'année j^{re} . Par conséquent, il est également raisonnable d'utiliser à sa place le poids médian :

(38)

$$w_{i,\text{moyen}}^j = \frac{1}{\pi_j^A} \text{ la médiane de } \pi_j^A, j = 1, 2, \dots, m^A.$$

pour accroître la robustesse des estimations. De même, nous estimons Y_B^j par

(39)

$$\hat{Y}_{B(1)}^j = \sum_{i=1}^n m_i^A \frac{L_{ji,i}}{w_{i,\text{moyen}}^j} \sum_{k=1}^K Y_{ik}^j,$$

et

(40)

$$\hat{Y}_{B(2)}^j = m_i^A \sum_{i=1}^n w_i^j \sum_{k=1}^K Y_{ik}^j.$$

Les comparaisons de ces méthodes proposées avec et sans intégration du problème de non-réponse en utilisant le poids moyen ainsi que le poids médian dans chaque ménage sont présentées aux tableaux 2 à 5.

Les quatre tableaux suivants donnent l'évaluation de la performance de nos estimations fondées sur l'erreur relative définie comme étant :

$$\left| \frac{\text{estimation} - \text{«valeur réelle»}}{\text{«valeur réelle»}} \right| \times 100 \, \%.$$

sections 3.1.1 et 3.1.2. Afin d'évaluer les propriétés des

estimations obtenues par ces approches, nous exécutons une étude par simulation en nous servant des données de l'EDTR. Nous nous intéressons aux estimations transversales pour quatre variables de revenu pour l'année 2003. Ces quatre variables sont le revenu total avant impôt, le revenu total après impôt, les gains (c'est-à-dire les traitements et salaires avant retenues et le revenu d'un travail autonome) et les salaires et traitements avant retenues (également appelés revenu d'emploi). Nous nous intéressons au total de population pour ces quatre variables. Nous avons estimé ces quatre grandeurs d'intérêt aux niveaux national et

provincial.

Dans une enquête longitudinale, le nombre total de liens dans la grappe i n'est généralement pas supérieur au nombre total d'individus dans cette grappe ni inférieur au nombre d'individus longitudinaux dans cette grappe. Puisque T_B^j est inconnu, nous remplaçons T_B^j par M_B^j dans (5) pour notre étude par simulation.

Premièrement, nous supposons que les liens entre toutes les unités sélectionnées durant l'année initiale (1999) et toutes les unités présentes dans l'ensemble de la population en 2003 sont spécifiées correctement. Puis, nous calculons les totaux en utilisant la MGPP. Nous utilisons ces totaux comme estimations cibles, c'est-à-dire les valeurs « réelles. »

Deuxièmement, nous supprimons aléatoirement 50 % des liens associés aux individus initialement présents en convertissant aléatoirement certains cohabitants initialement présents en cohabitants initialement absents. Le nombre de liens enlevés de la sorte représente environ 6,3 % de la population totale à laquelle nous nous intéressons, dont la taille est de 30 224. Sans aucune correction, nous recalculons l'estimation en utilisant la MGPP. Nous utilisons cette estimation comme estimation de référence c'est-à-dire le « placebo ».

Troisièmement, nous estimons les mêmes quantités en utilisant la MGPP avec les méthodes d'ajustement proportionnelles, c'est-à-dire les méthodes (1) et (2) de la section 3.1, pour voir si les estimations sont suffisamment proches de la « valeur réelle » et déterminer l'importance de l'amélioration due à ces ajustements.

Cette étude par simulation en utilisant les données de l'EDTR démontre que la méthode proposée donne de très bons résultats en ce qui concerne la correction de la surestimation due à la non-réponse de lien.

en utilisant les liens observés et leurs variables caractéristiques correspondantes. Les vecteurs de paramètres \mathbf{a} et \mathbf{b} peuvent être estimés. Nous proposons d'imputer les liens f_{ijk} par leur probabilité estimée :

$$\hat{p}_{f,ik} = \frac{1 + e^{\mathbf{a}\mathbf{x}_i' + \mathbf{b}\mathbf{x}_k'}}{e^{\mathbf{a}\mathbf{x}_i' + \mathbf{b}\mathbf{x}_k'} + 1} \quad (31)$$

où (\mathbf{a}, \mathbf{b}) est un estimateur de (\mathbf{a}, \mathbf{b}) ; par exemple, nous utilisons l'estimateur du maximum de vraisemblance (pseudo-vraisemblance) pondéré. Nous obtenons alors

$$\hat{L}_{B(3)}' = \sum_{j \in \Omega^A \setminus (i^A \cup \Delta_0^A)} L_{f,ij} + \sum_{M \in \Omega^A} \frac{1 + e^{\mathbf{a}\mathbf{x}_i' + \mathbf{b}\mathbf{x}_k'}}{e^{\mathbf{a}\mathbf{x}_i' + \mathbf{b}\mathbf{x}_k'} + 1} L_{f,ij} \quad (32)$$

En remplaçant $L_{B(3)}'$ par $L_{B(3)}$, (5) nous donne un estimateur convergent de X_B quand le modèle spécifié dans (30) est correct et que (\mathbf{a}, \mathbf{b}) est convergent. Notons qu'il existe d'autres options que le modèle logistique, telles que les modèles logit et log-log complémentaires. Pour plus de renseignements, consulter Draper et Smith (1998). Ces auteurs mentionnent aussi que le choix du modèle n'est pas toujours évident en pratique.

3.2.2 Estimation directe de L_B' en utilisant un modèle log-linéaire

Nous considérons qu'il existe un vecteur de variables \mathbf{x}_B' total de liens dans une grappe varie seulement en fonction des caractéristiques de la grappe proprement dite. En utilisant le modèle log-linéaire, nous pouvons proposer l'expression suivante :

$$\log(L_B') = \theta^T \mathbf{x}_B' \quad (33)$$

Si la qualité de l'ajustement est raisonnable, nous pouvons estimer L_B' directement par

$$L_{B(4)}' = e^{\theta^T \mathbf{x}_B'}, \quad (34)$$

où θ est un estimateur de θ . Si θ est convergent, après remplacement de L_B' par $L_{B(4)}'$, (5) nous donne un estimateur convergent de X_B . Nous notons que $L_{B(4)}'$ pourrait avoir une valeur non entière et donc devoir être arrondi à la valeur entière la plus proche.

4. Étude par simulation

Si, sous un plan de sondage longitudinal, on souhaite produire des estimations transversales à un point particulier dans le temps après le point de départ, la situation devient

Dans cet exemple particulier, U^A est la population à l'année initiale, disons y_0 , de l'enquête longitudinale et U^B est la population durant n importe quelle année, disons l'année y_1 , après l'année initiale. L'échantillon s^A est une forme de tous les individus longitudinaux. $L_{f,1}$ est une variable binaire; sa valeur est 1 si l'individu j vit dans le i^e ménage à l'année y_1 et elle est 0 autrement. L_B' est le nombre total de personnes longitudinales et de cohabitants initialement présents à l'année y_0 qui vivent dans le i^e ménage à l'année y_1 .

Le plan d'échantillonnage de l'EDTR est décrit en détail dans Lavallée (1993). Certains termes que nous utilisons dans le présent article, tels que cohabitants, individus initialement présents et individus initialement absents, sont ceux utilisés dans Lavallée (1995). Les individus initialement absents dans la population sont ceux qui ne faisaient pas partie de la population durant l'année où l'échantillon longitudinal a été sélection, mais qui sont considérés dans l'échantillon ultérieur; font partie de ce groupe les nouveaux et les immigrants. Après l'année initiale de sélection, la population compte des individus longitudinaux, des individus initialement présents et des individus initialement absents. Si nous nous concentrons sur les ménages contenant au moins un individu longitudinal (c'est-à-dire les ménages longitudinaux), les individus initialement présents ou absents qui se joignent à ces ménages sont appelés cohabitants.

Pour un individu longitudinal, le lien sera de type un à un (bivarié). Dans le cas des cohabitants, il est fort probable qu'il sera impossible d'identifier le lien quelques années après l'année initiale de l'enquête, à cause, par exemple, des nouvelles naissances et de l'immigration; en outre, plus la proportion de cohabitants dans la population cible est élevée, plus cette probabilité devient forte. Par exemple, dans le panel 3 de l'EDTR, les cohabitants représentaient 7,8 % de 47 377 individus en l'an 2000, c'est-à-dire un an après l'année initiale. La proportion était passée à 13,87 % en l'an 2002 (trois ans plus tard) et à 15,22 % en 2003 (quatre ans plus tard). Nous voyons donc que l'on ne peut pas fermer les yeux sur les non-réponses de l'enquête. Comme nous disposons d'information observée, nous exécutons l'estimation de L_B' par les deux types d'ajustements proportionnels que nous avons proposés aux

$$L_B^{*} = \sum_{j=1}^{J^A} \frac{\pi_{j,j}^A}{L_B^{j,j}} \quad (24)$$

Revenons à l'exemple de la figure 2 avec deux non-réponses de lien qui surviennent entre l'unité $j = 3$ dans U^A et les unités $k = 1, 2$ de la grappe $i = 2$ dans U^B . Afin d'appliquer (28), nous calculons d'abord m^A/T^A . Pour notre exemple, nous avons $m^A = 2$ et $T^A = 3$. L'estimateur résultant de Y^B en utilisant la méthode d'ajustement (2) pour cet exemple est alors

$$Y^B = \frac{3}{2} \left[\frac{1}{2} \left(\frac{\pi_1^A}{1} + \frac{\pi_2^A}{1} \right) Y_{11} + \frac{1}{2} \left(\frac{\pi_1^A}{1} + \frac{\pi_2^A}{1} \right) Y_{12} \right] + \frac{1}{1} Y_{21} + \frac{1}{1} Y_{22} + \frac{1}{1} Y_{23}. \quad (29)$$

Par conséquent, la correction faite dans (28) est différente de celle résultant de la méthode (1) dans cet exemple.

Nous savons que $\text{var}(Y_{B(1 \text{ ou } 2)} | s^A) = \text{var}\{E(Y_{B(1 \text{ ou } 2)} | s^A)\} + E\{\text{var}(Y_{B(1 \text{ ou } 2)} | s^A)\}$. L'espérance et la variance intégrées (conditionnellement à s^A) sont calculées sur tous les ensembles possibles de liens $I_{j,k}^B$ « répondants » sachant l'échantillon s^A , tandis que l'espérance et la variance extérieures sont calculées sur tous les échantillons s^A possibles. En général, les corrections faites ci-dessus n'éliminent pas le deuxième terme qui dépend du caractère aléatoire de $I_{j,k}^B$.

3.2 Estimation de L_B^i en disposant de variables auxiliaires

3.2.1 Estimation de $I_{j,k}^B$ en utilisant un modèle logistique

Les méthodes d'estimation de L_B^i proposées à la section 3.1 sont faciles à appliquer et ne nécessitent pas d'information supplémentaire. Toutefois, les hypothèses peuvent être violées, ce qui produit une estimation indésirable. Par exemple, L_B^i peut dépendre de certaines caractéristiques de l'unité j et de la grappe i .

Nous supposons que la probabilité qu'il existe un lien entre une unité dans la population d'échantillonnage et une unité dans la population cible dépend de certaines variables auxiliaires en spécifiant un modèle de régression logistique. Nous pouvons estimer cette fonction de probabilité de façon que l'estimation de la quantité d'intérêt dans la population cible soit désirable. Soit $P_{j,k}^i = P(I_{j,k}^i = 1)$ qui est affectée par un certain vecteur de variables \mathbf{x}_j^A dans U^A et \mathbf{x}_k^B dans U^B .

Nous pouvons ajuster le modèle logistique

$$\log \left(\frac{1 - P_{j,k}^i}{P_{j,k}^i} \right) = \mathbf{a}' \mathbf{x}_j^A + \mathbf{b}' \mathbf{x}_k^B \quad (30)$$

Donc, les poids d'estimation sont donnés par

$$\frac{m^A}{\sum_{j=1}^{J^A} L_B^{j,i}} = \frac{T^A}{\sum_{j=1}^{J^A} L_B^{j,i}}. \quad (26)$$

(A) Supposons que pour toute grappe i , la moyenne de liens existants associés avec toutes les unités comprises dans l'échantillon s^A est la même que celle des liens existants associés à toutes les unités comprises dans U^A , c'est-à-dire

$$L_B^{i(2)} = \frac{m^A}{T^A} \sum_{j=1}^{J^A} L_B^{j,i}. \quad (25)$$

Dans le cas de plans de sondage aléatoires simples avec ou sans stratification, $L_B^{j(2)}$ donne un estimateur sans biais de L_B^i . Pour les plans plus complexes, il fournit un estimateur fondé sur un modèle sans biais sous l'hypothèse (A)

de la façon suivante :

La deuxième approche que nous proposons ici consiste à estimer $\pi_{j,i}^A$ en utilisant la proportion d'unités dans s^A qui appartiennent à Ω^A , c'est-à-dire $\pi_{j,i}^A = m^A/T^A$. Cela nous informe que

une estimation grossière de la valeur. forme générale, mais nous pouvons habituellement donner un calcul est compliqué et varie d'un cas à l'autre sans avoir de complexité des effets de s^A sur Ω^B , donc sur Ω^A . Le il s'agit d'une fonction de π_j^A qui dépend pourtant de la de Ω^A . Nous devons maintenant calculer π_j^A . De nouveau, variable indicatrice de la présence dans s^A en provenance estimer l'information sur les liens dans T^A , où I_j^A est la nous servant de l'information sur les liens dans s^A pour

$$L_B^{*} = \sum_{j=1}^{J^A} \frac{\pi_{j,j}^A}{L_B^{j,j}}$$

dans la grappe i .

Il s'ensuit que Y^B peut être estimé par

$$Y_{B(2)}^i = \frac{T^A}{m^A} \sum_{j=1}^{J^A} \frac{\pi_{j,j}^A}{L_B^{j,i}} = \sum_{k=1}^{M^B} w_{j(2)}^i Y_{j,k}^B = \sum_{k=1}^{M^B} w_{j(2)}^i \sum_{j=1}^{J^A} \frac{\pi_{j,j}^A}{L_B^{j,i}} \quad (28)$$

$$\sum_n \sum_{M_B} \frac{\sum_{j=1}^{M_A} L_{j,i}^B}{t_j^A \pi_j^A} Y_{ik}^n = \sum_n \sum_{M_B} w_{(i)}^{(n)} Y_{ik}^n \quad (20)$$
$$(6I) \quad \sum_{v=1}^{l_f} \frac{1}{v^f L} = \frac{(1) \sum_{v=1}^{l_f} \frac{1}{v^f}}{1^f \sum_{v=1}^{l_f} \frac{1}{v^f}} = (1) \sum_{v=1}^{l_f} \frac{1}{v^f}$$
$$(8I) \quad \frac{\partial L}{\partial u} = \frac{\partial y}{\partial T}$$
$$(L1) \quad \left(\sum_{\substack{M \\ \#}}^{[=Y]} \frac{\sum_{\substack{N \\ \#}}^f \frac{u}{f} T}{\sum_{\substack{I \\ \#}}^f \frac{gT}{I}} \right)_{\substack{E \\ \#}} = \left(\sum_{\substack{u \\ \#}}^{[=I]} \frac{1}{u} \right)_{\substack{E \\ \#}} =$$
$$\left[E_{l'_i} \sum_{n=1}^i E_{l'_i} \left(\frac{1}{z} T_g^{(l'_i)} \right) 2l'_i T_g - \sum_{l'_i=1}^j \frac{\gamma_{l'_i}^u}{\gamma_{l'_i}^{l'_i}} T_g^{l'_i} \right] \approx$$
[illegible]
$$w_{(1)}^i = \frac{\sum_{j=1}^m L_B^{j,i} T_A^{j,i}}{\sum_{j=1}^m \pi_A^{j,i} L_B^{j,i}}.$$

(81)

[illegible]

(23)

manquants avec U_B . Soit $\Delta^s = \Omega^s \setminus s^s$ l'ensemble d'unités pour lesquelles des liens pourraient manquer. Alors,

$$(8) \quad L_B^t = \sum_{M_B^t} \sum_{k=1}^{j \in \Delta^s} I_{j,ik} + \sum_{M_B^t} \sum_{k=1}^{j \in \Delta^s} I_{j,ik}.$$

Si nous exécutons la MGPP sans tenir compte de ces liens manquants, nous serons du total des $I_{j,ik}$ observés comme valeur de L_B^{t*} pour calculer Y_B^t en utilisant

$$(9) \quad L_B^t = \sum_{M_B^t} \sum_{k=1}^{j \in \Delta^s} I_{j,ik} + \sum_{M_B^t} \sum_{k=1}^{j \in \Delta^s} I_{j,ik},$$

où Δ^s est un sous-ensemble de Δ^s qui contient uniquement les unités dont les liens sont observés. Le prix de cette approche est la surestimation de Y_B^t en utilisant (5),

puisque

$$L_B^t \geq L_B^{t*}.$$

Nous proposons quelques méthodes pour appliquer la MGPP en tenant compte de la non-réponse de lien dans l'estimation de L_B^t .

3.1 Estimation de L_B^t en l'absence de variables

auxiliaires

3.1.1 Estimation de L_B^t par ajustement proportionnel pour chaque grappe (méthode 1)

Pour aborder le problème de la non-réponse de lien, nous nous concentrons sur l'estimation de L_B^t en utilisant l'information connue sur les liens à l'intérieur de s^s . Pour calculer les poids donnés par (6) en utilisant la MGPP, il nous suffit d'estimer L_B^t pour les $i \in \Omega_B^t$. Pour tout $i \in \Omega_B^t$,

$$(10) \quad L_B^t = \sum_{T^s} L_B^{t,j,i},$$

Un estimateur général de ce total peut s'exprimer sous la forme

$$(11) \quad L_B^t = \sum_{T^s} w_{L_B^{t,j,i}}^s,$$

où $w_{L_B^{t,j,i}}^s$ est un poids aléatoire qui prend la valeur $w_{L_B^{t,j,i}}^s = 0$ si j n'est pas dans l'échantillon s^s . Pour chaque $i \in \Omega_B^t$, nous utilisons l'information connue sur les liens entre s^s et U_B^t pour estimer l'information sur les liens entre Ω_A^t et U_B^t . L'espérance de L_B^t est

$$(12) \quad E(L_B^t) = \sum_{T^s} E(w_{L_B^{t,j,i}}^s) L_B^{t,j,i}.$$

En comparant (10) et (12), nous constatons que L_B^t est sans biais pour L_B^t sous tout scénario de pondération avec $E(w_{L_B^{t,j,i}}^s) = 1$ pour tout j .
Thompson (Horvitz et Thompson 1952), également appelé estimateur π (Särndal, Swensson et Wretman 1991). Notons qu'en vertu de la définition de Ω_A^t , $\Omega_A^t \subset s^s$ pour tout i . Nous imitons une procédure d'estimation du nombre de liens dans Ω_A^t en utilisant celui dans s^s . La procédure consiste à sélectionner un « échantillon » s_A^s dans la « population » Ω_A^t . Soit $\pi_{L_B^{t,j,i}}^s$ la probabilité que j (qui est dans Ω_A^t) soit incluse dans s_A^s . Posons alors que

$$(13) \quad w_{L_B^{t,j,i}}^s = \begin{cases} 1/\pi_{L_B^{t,j,i}}^s, & j \text{ est dans } s_A^s, \\ 0, & j \text{ est dans } \Omega_A^t \setminus s_A^s. \end{cases}$$

Selon le corollaire 3.1 dans Casse, Särndal et Wretman (1977), ce scénario de pondération donne un estimateur sans biais de L_B^t . Nous avons

$$(14) \quad L_B^t = \sum_{T^s} \frac{w_{L_B^{t,j,i}}^s}{L_B^{t,j,i}} \pi_{L_B^{t,j,i}}^s.$$

Cela nous donne un estimateur asymptotiquement sans biais (la preuve suit) de Y_B^t :

$$(15) \quad Y_B^t = \sum_n \sum_{j=1}^{T^s} \frac{L_B^{t,j,i}}{L_B^{t,j,i}} \frac{\pi_{L_B^{t,j,i}}^s}{\sum_{k=1}^{M_B^t} Y_{B,k}^s}.$$

Afin de démontrer l'absence de biais de cet estimateur, nous employons un développement en série de Taylor. Selon le corollaire 5.1.5 (Fuller 1996), nous obtenons

$$\frac{1}{1} = \frac{1}{1} \left(\frac{L_B^t}{L_B^t} - \frac{L_B^t}{L_B^t} \right) + O_p(L_B^t - L_B^t),$$

$$\frac{1}{1} = \frac{1}{1} (2L_B^t - L_B^t) + O_p(n^{-1}).$$

Il s'ensuit que

$$p \lim \left\{ n \left[\frac{1}{2} \left(\frac{L_B^t}{L_B^t} - \frac{L_B^t}{L_B^t} \right) \right] \right\} = 0.$$

Par conséquent, en vertu du théorème 5.2.1 (Fuller 1996), la loi limite de $n^{1/2} [1/L_B^t - L_B^t/L_B^t]$ est la loi limite de $n^{1/2} [1/L_B^t - L_B^t/L_B^t]$. Nous notons que Y_B^t est une fonction de la variable aléatoire $t_{j,i}$ ainsi que de la variable aléatoire $t_{j,i}^s$; par conséquent, nous désignons l'espérance de Y_B^t par rapport à $t_{j,i}$ par $E_{t_{j,i}}(\cdot)$ et celle par rapport à $t_{j,i}^s$ par $E_{t_{j,i}^s}(\cdot)$. D'où nous obtenons asymptotiquement

Souignons que, dans le cas de l'estimateur pour lequel $l_{j,k}$ est connu, la seule hypothèse pour l'absence de biais est que $L_{j,i}^t > 0$ pour toutes les grappes i dans U^B . Autrement dit, chaque grappe de la population cible doit posséder au moins un lien provenant de U^A . Nous savons que si certains liens manquent, l'estimateur (5) sera biaisé. En cas de non-réponse de lien, comme il est mentionné dans Lavallée (2001), on ne peut déterminer $L_{j,i}^t$. Habituellement, l'utilisation du nombre total de liens observés pour remplacer cette quantité inconnue produit une surestimation de Y^B parce que certaines composantes des liens manquent réellement dans la sommation $L_{j,i}^t$. Ce problème est exactement celui que nous nous proposons d'examiner dans la présente étude et nous essayons de corriger les poids d'estimation w_{ik}^t en estimant $L_{j,i}^t$ de façon à obtenir une meilleure estimation de Y^B .

(3)

$$L_{j,i}^t = \sum_{k=1}^K \sum_{M^A} l_{j,ik}^t;$$

Étape 3 : Obtenir le poids final w_i

(4)

$$w_i = \frac{L_{j,i}^t}{\sum_{M^B} w_{ik}^t};$$

Étape 4 : Poser que $w_{ik} = w_i$ pour tout k dans la i^{e} grappe.

Il découle du théorème de la section 3 de Lavallée (2001) que

(5)

$$\hat{Y}^B = \sum_n \sum_{j=1}^J \frac{L_{j,i}^t}{\pi_{j,i}^A} \sum_{M^B} Y_{ik}^B$$

offre un estimateur sans biais de Y^B à condition que tous les liens $l_{j,ik}^t$ puissent être identifiés correctement. Les poids d'estimation attribués dans (5) sont

$$w_{ik} =$$

$$\left\{ \begin{array}{l} \sum_{j=1}^J \frac{L_{j,i}^t}{\pi_{j,i}^A} \text{ pour toutes les unités } k \text{ dans la} \\ \text{grappe } i \text{ quand } i \text{ est dans } \Omega^B; \\ 0, \text{ quand } i \text{ n'est pas dans } \Omega^B. \end{array} \right. \quad (6)$$

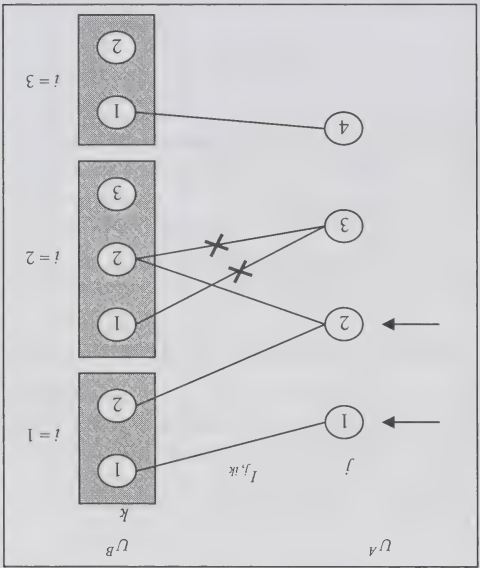
Un exemple simple est illustré à la figure 2. Nous

voulons estimer le total Y^B lié à la population cible U^B . Supposons que nous sélectionnons les unités $j = 1$ et 2 dans U^A . En sélectionnant l'unité $j = 1$, nous étudions les unités de la grappe $i = 1$. De même, en sélectionnant l'unité $j = 2$, nous étudions les unités des grappes $i = 1$, et 2. Par conséquent, nous avons $\Omega^B = \{1, 2\}$. Pour chaque unité k des grappes i de Ω^B , nous calculons les poids initiaux w_{ik}^t donnés par (2), le nombre total de liens qui existent entre les unités de U^A et les unités de U^B , $L_{j,i}^t$, et les poids finaux w_{ik}^t . Alors, d'après (5), l'estimateur résultant de Y^B prend la forme qui suit (voir Lavallée 2007, pages 17 et 18 pour plus de précisions) :

$$\hat{Y}^B = \left[\frac{2}{1} \pi_{j,i}^A + \frac{1}{1} \pi_{j,i}^A \right] Y_{11}^B + \left[\frac{2}{1} \pi_{j,i}^A + \frac{1}{1} \pi_{j,i}^A \right] Y_{12}^B + \left[\frac{2}{1} \pi_{j,i}^A + \frac{1}{1} \pi_{j,i}^A \right] Y_{21}^B + \left[\frac{2}{1} \pi_{j,i}^A + \frac{1}{1} \pi_{j,i}^A \right] Y_{22}^B + \frac{3\pi_{j,i}^A}{1} Y_{23}^B. \quad (7)$$

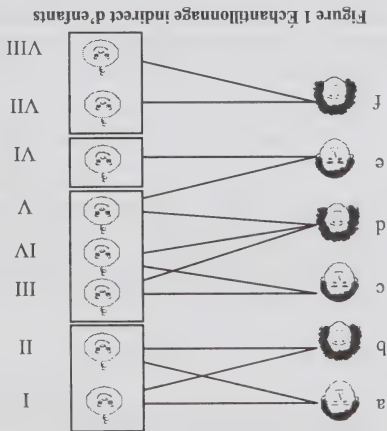
3. Traitements du biais dans les estimations

Figure 2 Exemple de liens dans l'échantillonnage indirect



Lavallée (2002), Deville et Lavallée (2006), et Lavallée

(2007).



2. Notation et problème

Soit U^A et U^B la population d'échantillonnage et la population cible, respectivement. Donc, U^A est la population relée à U^B pour laquelle existe une base de sondage connue. Soit s^A, M^A et m^A , un échantillon tiré de U^A , le nombre d'unités dans U^A et le nombre d'unités dans s^A , respectivement. Soit π_j^A la probabilité de sélection de la j^{e} unité dans U^A , avec $\pi_j^A > 0$ et $\sum_{j=1}^{M^A} \pi_j^A = m^A$. Nous utilisons également la notation suivante : M^B, N, U^B et M^B , le nombre d'unités dans U^B , le nombre de grappes dans U^B , la i^{e} grappe de U^B avec $\bigcup_{i=1}^N U_i^B = U^B$, et le nombre d'unités dans la i^{e} grappe U_i^B , respectivement. Soit $I_{j,k}^A$, une variable indicatrice de l'existence d'un lien : $I_{j,k}^A = 1$ indique qu'il existe un lien entre la j^{e} unité de U^A et la k^{e} unité de U^B , tandis que $I_{j,k}^A = 0$ indique qu'il n'existe pas de lien. Soit aussi L_j^B , le nombre total de liens existants entre l'unité j de U^A et les unités de U^B , c'est-à-dire $L_j^B = \sum_{k=1}^{M^B} I_{j,k}^A$. Soit L_i^B , le nombre total de liens existants entre les unités de U_i^A et les unités de U_i^B , c'est-à-dire $L_i^B = \sum_{j=1}^{M^A} L_j^B$. Nous désignons par $Y_{i,k}^B$ les caractéristiques de la k^{e} unité de la i^{e} grappe dans la population U^B , et par $Y_{i,k}^A$ le total de tous les $Y_{i,k}^A$. Nous avons alors $Y^B = \sum_{i=1}^N \sum_{k=1}^{M^B} Y_{i,k}^B$.

Désignons par Ω^B les grappes dans U^B où il existe au moins une unité i,k telle que $I_{i,k}^A = 1$ pour une j^{e} unité dans s^A , et disons qu'elle peut être identifiée par les unités j dans s^A , c'est-à-dire qu'une telle unité i satisfait $L_j^B = \sum_{k=1}^{M^B} I_{j,k}^A > 0$. Le nombre de grappes dans Ω^B est n . Après l'échantillonnage, nous avons réétiqueté les grappes comprises dans Ω^B au moyen de l'indice $i = 1, 2, \dots, n$. Nous désignons par w_k^A le poids d'estimation appliqué à la

$$Y^B = \sum_{i=1}^n \sum_{k=1}^{M^B} w_{i,k}^A Y_{i,k}^B \quad (1)$$

En appliquant la MGPP, nous attribuons un poids d'estimation w_k^A à chaque unité k des grappes étudiées. Ces poids peuvent être choisis de manière appropriée pour que l'estimateur de Y^B :

$$f_{j,k}^B = \frac{Y_{j,k}^B}{w_{j,k}^A} \quad (2)$$

calculant les poids pour chaque grappe qui a été observée.

Étape 1 : Fournir les poids initiaux w_k^A .

$f_{j,k}^B = 1$ indique le contraire.

Notre objectif est d'estimer le total Y^B , qui est notre paramètre d'intérêt, pour la population cible U^B qui est divisée en N grappes. Pour le faire, nous sélectionnons un échantillon s^A dans U^A avec la probabilité de sélection π_j^A . Puis, nous identifions Ω^B en utilisant $f_{j,k}^B \neq 0$. Toutes les unités des grappes comprises dans Ω^B font partie de l'étude dans laquelle nous mesurons $Y_{i,k}^B$ et l'ensemble des unités comprises dans U^A qui ont des liens avec certaines unités de U^B avec $i \in \Omega^B$, et par Ω^A l'ensemble d'unités comprises dans U^A qui ont des liens avec certaines unités de U^B avec $i \in \Omega^B$. Soit T^A, T^A et m^A , le nombre d'unités dans Ω^A , le nombre d'unités dans s^A , et le nombre d'unités dans s^A , respectivement. Enfin, nous utilisons les trois indicateurs suivants : soit f_j la variable indicatrice de sélection dans s^A ; $f_j = 1$ indique que la j^{e} unité de U^A est dans s^A et $f_j = 0$ indique le contraire. Soit f_j^A la variable indicatrice d'inclusion dans s^A pour les unités de Ω^A ; $f_j^A = 1$ indique que la j^{e} unité de Ω^A est dans s^A et $f_j^A = 0$ indique le contraire, soit, enfin, $f_{j,i}^A$ la variable indicatrice d'inclusion dans s^A pour les unités de Ω^A ; $f_{j,i}^A = 1$ indique que la j^{e} unité de Ω^A est dans s^A et $f_{j,i}^A = 0$ indique le contraire.

Traitements de la non-réponse de lien dans l'échantillonnage indirect

Xiaojian Xu et Pierre Lavallée¹

Résumé

Nous cherchons à corriger la surestimation causée par la non-réponse de lien dans l'échantillonnage indirect lorsque l'on utilise la méthode généralisée de partage des poids (MGPP). Nous avons élaboré quelques méthodes de correction pour tenir compte de la non-réponse de lien dans la MGPP applicables lorsque l'on dispose ou non de variables auxiliaires. Nous présentons une étude par simulation de certains de ces méthodes de correction fondée sur des données d'enquête longitudinale. Les résultats des simulations révèlent que les corrections proposées de la MGPP réduisent bien le biais et la variance d'estimation. L'accroissement de la réduction du biais est significatif.

Mots-clés : Méthode de partage des poids ; non-réponse ; échantillonnage indirect ; enquête longitudinale.

1. Introduction

Par échantillonnage indirect, on entend la sélection d'échantillons dans une population qui n'est pas celle que l'on veut étudier, mais qui y est reliée. Un scénario d'échantillonnage de ce genre est souvent mis en œuvre lorsque l'on ne dispose pas de bases de sondage pour la population cible, mais que l'on en possède pour une autre population qui y est reliée. Nous appelons cette dernière la population d'échantillonnage. Par exemple, dans Lavallée (2007), nous considérons la situation où l'estimation a trait aux jeunes enfants dans les familles, mais où nous disposons seulement d'une liste de noms de parents comme base de sondage. Par conséquent, nous devons d'abord sélectionner un échantillon de parents avant de pouvoir sélectionner l'échantillon d'enfants. Dans cette situation type d'échantillonnage indirect, la population d'échantillonnage est celle des parents, tandis que la population cible est celle des enfants. Il convient de souligner que les enfants d'une famille particulière peuvent être sélectionnés par l'entremise de leur père ou de leur mère. La figure 1 donne une illustration simple de ce scénario d'échantillonnage indirect (figure 1.2, Lavallée 2007).

La littérature concernant les problèmes d'estimation associés à l'échantillonnage indirect est abondante et nous en nommerons que quelques-uns de ces travaux ici. Ernst (1989) est le premier à discuter des méthodes appliquées pour produire des estimations transversales au moyen de données provenant d'une enquête-ménages longitudinale. Il présente la méthode de partage des poids dans le contexte d'une enquête longitudinale et montre aussi qu'elle fournit un estimateur sans biais du total pour tout caractèreistique de la population d'intérêt. Kalton et Brick (1995) concluent qu'une telle méthode fournit aussi la variance minimale du total de population estimée sous certains scénarios d'échantillonnage simples pour une enquête-ménages longitudinale

par panel. Lavallée (1995) étend la méthode de partage des poids à un contexte tout à fait général d'échantillonnage indirect qui comprend l'enquête longitudinale comme cas particulier, pour obtenir la méthode généralisée de partage des poids (MGPP). Il montre que ce scénario de pondération produit des estimations sans biais quel que soit le plan d'échantillonnage utilisé pour obtenir un échantillon de la population d'échantillonnage. Comme cela est le cas de tout scénario de pondération, la mise en œuvre de la MGPP requiert une correction pour divers problèmes de non-réponse. Lavallée (2001) fournit une MGPP corrigée tenant compte des problèmes de non-réponse dans l'échantillonnage indirect. Ce dernier comporte un autre type de non-réponse, dénommée non-réponse de lien (*link nonresponse*) et que Lavallée (2001) avait appelée « non-réponse de relation (*relationship nonresponse*) », dû au fait qu'il est impossible de déterminer, ou que l'on n'a pas réussi à déterminer, si une unité de la population d'échantillonnage est apparentée (liée) à une unité dans la population cible. Lavallée (2001) souligne qu'en cas de non-réponse de lien, la MGPP donne lieu à une surestimation, mais laisse ouverte la question de la correction qu'il convient d'apporter pour tenir compte de la non-réponse de lien. L'objectif de la présente étude est d'élaborer des méthodes pour traiter le biais d'estimation causé par cette non-réponse de lien.

La présentation de la suite de l'article est la suivante. À la section 2, nous décrivons la notation et le problème étudié. À la section 3, nous proposons quelques modifications de la MGPP en vue d'intégrer la non-réponse de lien. À la section 4, nous décrivons une étude par simulation portant sur un ensemble de données réelles et à la section 5, nous présentons certaines conclusions. Il convient de souligner que nous nous servons dans le présent article d'une étude par simulation pour illustrer les progrès réalisés grâce aux nouvelles méthodes, et que d'autres contributions théoriques à la résolution du problème peuvent être consultées dans

1. Xiaojian Xu, Département de mathématique, Université Brock, St. Catharines (Ontario) Canada, L2S 3A1. Courriel : xxu@brocku.ca ; Pierre Lavallée, Division des méthodes et enquêtes sociales, Statistique Canada, Ottawa (Ontario), K1A 0T6. Courriel : Pierre.lavallee@statcan.gc.ca.

- Singh, R., Singh, S., Mangat, N.S. et Tracy, D.S. (1995). An improved two stage randomized response strategy. *Statistical Papers*, 36, 265-271.
- Singh, S., Horn, S., Singh, R. et Mangat, N.S. (2003). On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in Transition*, 6 (4), 515-522.
- Singh, S., Singh, R., Mangat, N.S. et Tracy, D.S. (1994). An alternative device for randomized responses. *Statistica*, 54, 233-243.
- Skinner, C., Marsh, C., Openshaw, S. et Wymer, C. (1994). *Statistics*, 10 (1), 31-51.
- Soeken, K.L., et Macready, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92 (2), 487-489.
- Tracy, D.S., et Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade – A follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4 (2/3), 147-158.
- van den Hout, A., et van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *Revue Internationale de Statistique*, 70 (2), 269-288.
- Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- Willenborg, L., et de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York : Springer.
- Winkler, W.E. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. *Research Report Series of the Statistical Research Division of the U.S. Bureau of the Census*, #2004-06.

Pour $V^R(x_i)$, nous avons

$$\frac{1}{V^R(x_i)} = \frac{v^R}{V^R(y_i)}$$

et

$$V^R(y_i) = b + a \cdot x_i - (b + a \cdot x_i)^2$$

$$= (b + a \cdot x_i) \cdot (1 - b - a \cdot x_i)$$

$$= b \cdot (1 - b) + a \cdot (1 - 2 \cdot b - a) \cdot x_i.$$

Alors

$$E_p(V^R(\pi_A | s)) =$$

$$\frac{1}{N^2} \cdot \left(b \cdot (1 - b) \cdot \sum_i \frac{a^2}{1} + \frac{a}{1 - 2 \cdot b - a} \cdot \sum_i \frac{a \cdot x_i}{1} \right).$$

Cela complète la preuve du théorème 2.

Bibliographie

- Chaudhuri, A., et Mukerjee, R. (1987). *Randomized Response*. New York : Marcel Dekker.
- Dalenius, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86.
- Dalenius, T., et Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.
- Defays, D., et Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14 (4), 449-461.
- Domingo-Ferrer, J., et Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.
- Fidler, D.S., et Kleinknecht, R.E. (1977). Randomized response versus direct questioning: Two data collection methods for sensitive information. *Psychological Bulletin*, 84 (5), 1045-1049.
- Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9 (2), 383-406.
- Gouweleuw, J.M., Kooiman, P., Willenborg, L.C.R.J. et de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14 (4), 463-478.
- Greenberg, B.G., Abul-El-A., Simmons, W.R. et Horvitz, D.G. (1969). The unrelated question randomized response model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Hoboken : John Wiley & Sons, Inc.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. Hoboken : John Wiley & Sons, Inc.
- Horvitz, D.G., Shah, B.V. et Simmons, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 65-72.
- Hua, M., et Pei, J. (2008). A survey of utility-based privacy-preserving data transformation methods. Dans : *Privacy-preserving Data Mining: Models and Algorithms*, (Eds., C.C. Aggarwal et P.S. Yu), New York : Springer, 207-238.
- Kim, J. (1987). A further development of the randomized response technique for masking dichotomous variables. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 239-244.
- Kim, J.M., et Warde, W.D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, 133, 211-221.
- Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77 (2), 436-438.
- Landshert, J.A., van der Heijden, P. et van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, 33, 1-12.
- Leysteffer, F.W., et Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- Mangat, N.S. (1992). Two stage randomized response sampling procedure using unrelated question. *Journal of the Indian Society of Agricultural Statistics*, 44, 82-87.
- Mangat, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society, Series B*, 56, 93-95.
- Mangat, N.S., et Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.
- Mangat, N.S., Singh, S. et Singh, R. (1993). On the use of a modified randomization device in randomized response inquiries. *Metrica*, 51, 211-216.
- Nathan, G. (1988). Bibliographie de la méthode des réponses randomisées : 1965-1987. *Techniques d'enquête*, 14, 351-365.
- Quatember, A. (2007). Comparing the efficiency of randomized response techniques under uniform conditions. *IFAS Research Paper Series*, 23, www.ifas.jku.at/e2550/e2756/index_ger.html.
- Rosenberg, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 311-316.
- Sämdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer.
- Scheeters, N.J., et Dayton, C.M. (1987). Improved estimation of academic cheating behaviour using the randomized response technique. *Research in Higher Education*, 26 (1), 61-69.
- Singer, E., Mathiowetz, N.A. et Couper, M.P. (1993). The impact of privacy and confidentiality concerns on survey participation. The case of the 1990 U.S. Census. *The Public Opinion Quarterly*, 57 (4), 465-482.
- Singer, E., van Hoewyk, J. et Neugebauer, R.J. (2003). Attitudes and participation in the 2000 Census. *The Public Opinion Quarterly*, 67 (3), 368-384.

stratégies le permettent. Pour certains plans, y compris la question sur l'appartenance à une sous-population sensible non reliée à l'attribut étudié, il est nécessaire de trouver une sous-population adéquate de taille relative prédéterminée. D'autres plans peuvent être appliqués à des sous-populations de toute taille et sont donc plus pratiques. Par conséquent, une personne qui veut recueillir ou publier des données pourrait choisir, parmi les plans d'interrogation ayant la même efficacité, celui qui semble pouvoir être appliqué plus facilement que les autres.

Remerciements

L'auteur remercie le rédacteur associé et deux examinateurs de leurs précieux commentaires et suggestions.

Annexe

Preuves des théorèmes 1 et 2

Preuve du théorème 1 :

$$E(\pi_A) = \frac{1}{N} \cdot E_p \left(E_r \left(\sum_{x_i} \pi_{x_i} \mid s \right) \right) \\ = \frac{1}{N} \cdot E_p \left(\sum_{x_i} \frac{\pi_{x_i}}{x_i} \right) = \frac{1}{N} \cdot \sum_{x_i} \pi_{x_i} = \pi_A$$

La variance de l'estimateur (5) est donnée par

$$V(\pi_A) = V_p(E_r(\pi_A \mid s)) + E_p(V_r(\pi_A \mid s)).$$

Ainsi

$$V_p(E_r(\pi_A \mid s)) = \frac{1}{N^2} \cdot V_p \left(\sum_{x_i} \pi_{x_i} \right).$$

Soit l'indicateur d'inclusion dans l'échantillon

$$I_i = \begin{cases} 1 & \text{si l'unité } i \in s, \\ 0 & \text{autrement.} \end{cases}$$

Comme la covariance $C_r(\hat{x}_i, \hat{x}_j \mid s) = 0 \forall i \neq j$, pour le deuxième terme de $V(\pi_A)$, s'applique

$$E_p \left(\frac{1}{N^2} \cdot V_p \left(\sum_{x_i} \pi_{x_i} \right) \mid s \right) = E_p \left(\frac{1}{N^2} \cdot V_r \left(\sum_{x_i} \pi_{x_i} \mid s \right) \right) \\ = E_p \left(\frac{1}{N^2} \cdot \sum_{x_i} \frac{\pi_{x_i}^2}{x_i^2} \cdot V_r(x_i) \right) \\ = \frac{1}{N^2} \cdot \sum_{x_i} V_r(x_i) \cdot \pi_{x_i}.$$

formule générale pour la variance de l'estimateur sous échantillonnage probabiliste. Différents plans d'interrogation, en partie publiés et en partie – autant que nous sachions – non publiés jusqu'à présent, peuvent être considérés comme des cas particuliers de la stratégie standardisée (voir le tableau 1). Afin de comparer l'efficacité de ces plans d'interrogation, il est essentiel de tenir compte du niveau de protection de la vie privée qu'ils offrent. L'utilisation, pour cela, des « mesures λ » de la perte de vie privée décrites à la section 3 brosse un tableau entièrement différent de celui donné par presque toutes les publications antérieures ou sensibles doivent être répartis en diverses catégories afin de trouver le plan d'interrogation à variance minimale pour un niveau donné de protection de la vie privée (voir le tableau 2). La première catégorie comprend les sujets qui n'ont aucun caractère sensible. La deuxième comprend les sujets pour lesquels la possession, mais non la non-possession, d'un certain attribut est embarrassante pour les enquêtes. La dernière catégorie regroupe les sujets qui, dans leur ensemble, ont un caractère sensible.

Pour les sujets qui rentrent dans la première catégorie, il est tout à fait clair qu'aucune stratégie ne peut être plus efficace que l'interrogation directe (ST1 du tableau 1). En ce qui concerne les sujets de la deuxième catégorie, il n'existe qu'un seul plan d'interrogation permettant d'obtenir la variance minimale de l'estimateur. Il s'agit du plan selon lequel chaque enquête doit, soit avec la probabilité p_1 , répondre à la question sur l'appartenance au groupe ayant l'attribut sensible, soit avec la probabilité $1 - p_1$, répondre « oui » (ST4). Tous les autres cas particuliers de la stratégie standardisée protègent la vie privée de la personne interrogée, non seulement dans le cas d'une réponse « oui » comme le fait le scénario ST4, mais aussi dans le cas d'une réponse « non ». Par conséquent, leur performance ne peut pas atteindre le niveau de variance minimal réalisable.

Pour les sujets appartenant à la troisième catégorie, nous montrons que, contrairement à ce qu'affirment les auteurs d'autres publications, il n'existe aucune autre stratégie dominant de meilleurs résultats que celle proposée par Warner en 1965, dans les conditions où l'appartenance au sous-groupe étudié a un caractère aussi sensible que l'appartenance au complètement de ce sous-groupe. De nombreux autres plans sont aussi efficaces que celui de Warner, mais aucun n'est plus efficace.

Pour les variables de cette catégorie où l'appartenance à un groupe a un caractère sensible, mais pas aussi sensible que l'appartenance au groupe complètement, la situation change radicalement. Comparée sous les mêmes niveaux de protection de la vie privée, la méthode de Warner ne permet plus d'atteindre le meilleur résultat réalisable du plan randomisé standardisé, tandis que de nombreuses autres

indiquant la même perte de vie privée pour une réponse « oui » pour les deux plans d'interrogation, la somme des nombres affichés sur les deux dés devait être comprise entre 3 et 9 pour appliquer un paramètre de plan $p_1 = 0,805$. Les mesures λ de perte de vie privée pour ce choix sont données par $\lambda_1 = \lambda_0 = 4,143$, indiquant une perte de vie privée un peu plus élevée que dans le cas du plan ST_4 . Avec une probabilité de 0,805, les étudiants devaient répondre à la question « Êtes-vous membre du groupe U_4 ? » et avec la probabilité restante, à la question alternative « Êtes-vous membre du groupe U_4^c ? ».

Dans ces conditions, 38 seulement des 80 étudiants ont répondu « oui », ce qui donne une proportion estimée de « tricheurs » de

$$\hat{\pi}_{ST_2}^{\lambda} = \frac{\hat{\pi}_{ST_2}^y - p_2}{p_1 - p_2} = \frac{0,194}{0,475 - 0,194} = \frac{0,61}{0,4590}.$$

En plus de ce léger accroissement de la perte objective de vie privée, il existe une autre explication raisonnable pour ce résultat nettement plus faible. Quoique λ_1 n'ait pas tellement varié, certaines personnes participant à l'expérience doivent avoir été irritées par l'accroissement de la valeur p_1 jusqu'à 0,805 après qu'on leur ait demandé pour ST_4 , si elles continueraient de coopérer si p_1 devenait plus élevée que 0,75. En n'étant plus capables de faire la distinction entre la perte de vie privée causée par divers paramètres de plans différents plans d'interrogation, certains « tricheurs » ne voulaient plus continuer de répondre sincèrement. Simplement pour illustrer l'effet de divers plans d'interrogation sur l'efficacité du processus d'estimation, nous calculons l'estimateur de la variance de $\hat{\pi}_{ST_2}^{\lambda}$:

$$\hat{V}(\hat{\pi}_{ST_2}^{\lambda}) = \frac{p_1 \cdot (1 - p_1)}{p_1 \cdot (2p_1 - 1)^2} \cdot u = \frac{5,243 \cdot 10^{-3}}{5,243 \cdot 10^{-3}}.$$

L'accroissement considérable de la variance estimée est dû au fait que la stratégie de Warner ne protège pas toujours une réponse « non » de la même manière qu'une réponse « oui ». Puisque, dans notre cas, une réponse « non » ne devait pas être protégée du tout, cette protection inutile a eu un prix en terme d'exactitude.

6. Sommaire

Les stratégies fondées sur la réponse aléatoire ont été élaborées au départ pour réduire les taux de non-réponses et de réponses mensongères aux questions sur des sujets sensibles dans les enquêtes par sondage, mais elles peuvent aussi être appliquées aux fichiers de microdonnées à grande diffusion comme méthodes de masquage. La standardisation de ces méthodes pour l'estimation de proportions exposée dans le présent article offre une occasion d'établir une

Avant le sondage, nous avons essayé d'expliquer les conséquences de cette stratégie de randomisation sur la protection de la vie privée. Après que les étudiants aient donné la réponse sur la première feuille du questionnaire, seules ces feuilles ont été recueillies. En tout, 63 des 80 étudiants ont répondu « oui ». Nous nous attendions à ce que 20 des 80 étudiants le fassent parce qu'ils avaient reçu l'instruction de dire oui ». Par conséquent, en principe, 43 des 60 autres étudiants ont répondu « oui » à la question sensible. L'estimateur de π_{λ} est donné par

$$\hat{\pi}_{ST_4}^{\lambda} = \frac{\hat{\pi}_{ST_4}^y - p_4}{p_1 - p_4} = \frac{0,75}{0,7875 - 0,25} = \frac{0,75}{0,716}.$$

Pour ce sondage de population, la variance estimée de

π_{λ} est alors

$$\hat{V}(\hat{\pi}_{ST_4}^{\lambda}) = \frac{1 - p_1}{p_1} \cdot u \cdot (1 - \pi_{ST_4}^{\lambda}) = 1,181 \cdot 10^{-3}.$$

Après avoir achevé ce plan d'interrogation, nous avons demandé directement aux étudiants d'indiquer sur la deuxième feuille du questionnaire s'ils avaient répondu sincèrement à la première question ou non. Quatre étudiants seulement ont répondu par la négative. Cela signifie, si cela est vrai, que quatre étudiants de plus ont sans doute effectivement triché. La question suivante à laquelle les étudiants devaient répondre était celle de savoir s'ils continueraient de coopérer si p_1 (de ST_4) était supérieure à 0,75. En tout, 32 des 80 étudiants ont accepté de poursuivre, mais les autres non. Manifestement (au moins) quatre d'entre eux n'ont pas coopéré quand p_1 était égale à 0,75.

Enfin, nous avons appliqué la technique de Warner avec la même question sensible que pour le plan ST_4 appliqué auparavant. Afin de nous rapprocher d'un niveau λ_1 de 4,

Tableau 2 Paramètres de plan optimaux pour λ_1 et λ_0 donnés et divers types de caractères sensibles de la variable étudiée

Plan d'interrogation (catégorie de sujets)		Paramètres de plan qui optimisent la variance
ST1	(C ₁)	$p_1 = 1$
ST2	(C ₄)	$p_1 = \frac{\lambda_1}{\lambda_1 + 1}, p_2 = 1 - p_1$
ST3	(C ₃ , C ₄)	$\pi_B = \frac{\lambda_0 - 1}{\lambda_0} p_1 = \frac{\lambda_1 + \lambda_0 - 1}{(\lambda_1 - 1)(\lambda_0 - 1)}, p_3 = 1 - p_1$
ST4	(C ₂)	$p_1 = \frac{\lambda_1}{\lambda_1 - 1}, p_4 = 1 - p_1$
ST6	(C ₄)	$\pi_B = 0,5, p_1 : \frac{\lambda_1}{\lambda_1 + 1}, p_2 = p_1 - \frac{\lambda_1}{\lambda_1 + 1}$
ST6	(C ₃)	$p_3 = 1 - p_1 - p_2, \pi_B : \frac{\lambda_1}{\lambda_0 - 1} > \pi_B > \frac{1}{\lambda_0 - 1}$
ST7	(C ₃)	$p_1 = \frac{\lambda_1 \lambda_0 - \lambda_0}{\lambda_1 - 1}, p_2 = \frac{\lambda_1 - 1}{\lambda_1}, p_4 = 1 - p_1 - p_2$
ST9	(C ₃ , C ₄)	$\pi_B : 0 < \pi_B > \frac{\lambda_1 + \lambda_0 - 1}{\lambda_0}, p_1 = \frac{\lambda_1 - 1}{\lambda_0}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}$
ST10	(C ₃ , C ₄)	$\pi_B : \frac{\lambda_1 + \lambda_0 - 1}{\lambda_0} > \pi_B > \frac{\lambda_1 - 1}{\lambda_0}, p_1 = \frac{\lambda_1 - 1}{\lambda_0}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}$
ST11	(C ₃ , C ₄)	$p_1 = \frac{\lambda_1}{\lambda_1 - 1}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}, p_4 = \frac{\lambda_1 - 1}{\lambda_0}$
ST12	(C ₃ , C ₄)	$p_1 = \frac{\lambda_1}{\lambda_1 - 1}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}, p_4 = \frac{\lambda_1 - 1}{\lambda_0}$
ST13	(C ₃ , C ₄)	$p_1 = \frac{\lambda_1}{\lambda_1 - 1}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}, p_4 = \frac{\lambda_1 - 1}{\lambda_0}$
ST14	(C ₃ , C ₄)	$p_1 = \frac{\lambda_1}{\lambda_1 - 1}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}, p_4 = \frac{\lambda_1 - 1}{\lambda_0}$
ST15	(C ₃ , C ₄)	$\pi_B : 0 > \pi_B > \frac{\lambda_1 - 1}{\lambda_0}, p_1 = \frac{\lambda_1 - 1}{\lambda_0}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}$
ST16	(C ₃ , C ₄)	$\pi_B : 0 > \pi_B > \frac{\lambda_1 - 1}{\lambda_0}, p_1 = \frac{\lambda_1 - 1}{\lambda_0}, p_2 = \frac{\lambda_1 - 1}{\lambda_0}, p_3 = \frac{\lambda_1 - 1}{\lambda_0}$

stratégie de Greenberg et de ses collaborateurs avec π_B connue ($ST3$) a, d'une part, l'avantage par rapport au plan de Warner de donner des résultats optimaux également si $\lambda_{1,opt} < \lambda_{0,opt}$. D'autre part, toutefois, il a l'inconvénient (comme $ST6$) que la taille π_B de la sous-population U_B est entièrement prédéterminée (ou du moins bornée par un intervalle) si nous voulons atteindre l'efficacité optimale. En pratique, cela signifie que nous devons trouver une sous-population non reliée à la possession et à la non-possession de l'attribut A et de taille relative appropriée pour pouvoir obtenir l'exactitude optimale de l'estimateur. En principe, en fait être utilisé. Enfin, les cas particuliers les plus complexes, $ST15$ et $ST16$, de notre stratégie standardisée de réponse aléatoire peuvent tous les deux être utilisés avec une sous-population $U_B \subset U$ pour obtenir les meilleurs résultats.

5. Un exemple fondé sur des données réelles

Nous avons exécuté une étude empirique afin d'illustrer l'application de la stratégie à un plan d'interrogation. À cette fin, la population formée des 80 étudiants qui étaient inscrits au cours de « Statistique II » donne par l'auteur à l'Université Johannes Kepler à Linz (Autriche) durant le semestre du printemps 2009 a participé volontairement à une enquête. Le sujet étudié était le comportement de tricherie des étudiants. Pour les besoins de l'étude, la tricherie a été définie comme tout comportement qui n'était pas permis durant les examens écrits (y compris simplement copier les réponses des autres étudiants ou utiliser des documents interdits). Il ne fait aucun doute que le sujet est sensible pour ce genre de population. De surcroît, durant l'enquête, nous les étudiants étaient assis dans une salle de cours. Le paramètre d'intérêt était la proportion de la population d'étudiants qui avaient triché durant au moins l'un des examens du semestre précédent (y compris l'examen du cours Statistique I donné par l'auteur). Par conséquent, nous pouvons supposer d'une manière quasiment certaine que l'interrogation directe sur le sujet aurait donné lieu à une sous-estimation importante de cette proportion. Par exemple, une étude empirique menée par Scheers et Dayton (1987) a révélé de très faibles proportions pour presque tous les comportements de tricherie qu'ils sont examinés quand les questions sur le sujet étaient posées directement. L'utilisation de la stratégie de réponse aléatoire $ST3$ de Greenberg entraînait un accroissement important de ces proportions (ibidem, page 68).

Si nous recherchons les valeurs des paramètres du plan pour lesquelles la stratégie standardisée de réponse aléatoire peut donner cette variance et pour lesquelles les équations (14) et (15) sont vérifiées, nous constatons que dans ce cas il n'existe qu'une seule solution. Le seul plan d'interrogation capable de donner des résultats optimaux est $ST4$. Les paramètres qui optimisent la variance sont donnés par $p_1 = (\lambda_1 - 1/\lambda_1)$ et $p_4 = 1 - p_1$ (voir le tableau 2). Autrement dit, avec la probabilité $p_1 = (\lambda_1 - 1)/\lambda_1$, on demande à l'enquête s'il fait partie du groupe U_A et avec la probabilité restante, on lui donne l'insinuation de dire « oui ». De cette façon, l'enquêteur ne peut tirer une conclusion que d'une réponse « non » directement à la question non sensible de non-possession de la caractéristique A , mais non d'une réponse « oui » à la question de la possession de cet attribut sensible ou identificateur.

Le plan d'interrogation $ST1$ n'est pas applicable à de tels sujets, parce qu'il ne protège pas du tout la vie privée de l'enquête dans le cas d'une réponse « oui ». Toutes les autres méthodes proposent une réponse « non » plus qu'il n'est nécessaire. Par conséquent, elles peuvent être utilisées, mais elles ne permettent pas d'atteindre le degré d'efficacité de l'option $ST4$.

Si l'appartenance aux groupes U_A ainsi que U_A^* est de caractère sensible, de sorte que la variable est sensible dans son ensemble (par exemple : $U_A^* =$ l'ensemble de personnes mariées qui ont eu au moins une relation sexuelle avec leur partenaire la semaine précédente ; la condition $U_A^* = U - U_A$, $\lambda_{1,opt} < \lambda_{0,opt} < \infty$ s'applique. Dans ce cas, ni l'interrogation directe sur le sujet ni le plan $ST4$ ne peut être utilisé, parce que ces options ne permettent pas de protéger les deux réponses possibles.

Les autres plans sont applicables dans ces conditions, mais le plan de Warner ne permet pas d'atteindre le degré d'efficacité des autres si $\lambda_{1,opt} < \lambda_{0,opt}$. Il en est ainsi parce que ce plan protège toujours la vie privée de l'enquête aussi bien dans le cas d'une réponse « oui » que d'une réponse « non ». Toutefois, si $\lambda_{1,opt} = \lambda_{0,opt}$ malgré les allégations faites dans certaines publications dans le passé (voir, par exemple, Greenberg et coll., 1969, page 526f, Mangat et Singh 1990, page 440, Singh et coll., 2002, page 518f), il n'existe aucune méthode de réponse aléatoire qui peut donner de meilleurs résultats que la méthode $ST2$ de Warner avec les paramètres de plan optimaux p_1 et p_2 conformément au tableau 2. Pour $ST7$, cela n'est valide que si $\lambda_{1,opt} < \lambda_{0,opt}$. Par conséquent, $ST7$ est le supplément parait de $ST2$, pour lequel la situation est tout à fait opposée.

Tous les autres plans du tableau 1, tels que $ST1$ ou $ST14$, peuvent avoir la même efficacité pour $\lambda_{1,opt} < \lambda_{0,opt} < \infty$ si les paramètres du plan sont choisis conformément aux contraintes (14) et (15). Parmi ces plans, la

stratégies qui donnent les meilleurs résultats. Nous répondons à cette question à la section 4. Mais pour cela, nous devons prendre en considération le niveau de protection de la vie privée, qui varie selon le choix de ces paramètres.

3. Protection de la vie privée

Afin de pouvoir comparer l'efficacité des plans d'interrogation caractérisés par des paramètres de plan différents, il paraît inévitable de mesurer la perte de vie privée induite par ces paramètres. Nous pouvons pour cela utiliser les ratios λ_1 et λ_0 des probabilités conditionnelles qui suivent (voir, par exemple, les « mesures de mise en péril » (*measures of jeopardy*) dans Leysterfieter et Warner 1976, page 650) :

$$\lambda_j = \frac{\max[P(y_i = j \mid i \in U_A), P(y_i = j \mid i \in U_A^c)]}{\min[P(y_i = j \mid i \in U_A), P(y_i = j \mid i \in U_A^c)]} \quad (10)$$

($1 \leq \lambda_j \leq \infty$; $j = 1, 0$). Pour $j = 1$, (10) fait référence à la protection de la vie privée par rapport à une réponse « oui » et pour $j = 0$, par rapport à une réponse « non ». Pour le plan d'interrogation standardisé, ces « mesures λ » de perte de vie privée sont données par

$$\lambda_1 = \frac{\max[a + b; b]}{\min[a + b; b]} \quad (11)$$

$$\lambda_0 = \frac{\max[1 - (a + b); 1 - b]}{\min[1 - (a + b); 1 - b]} \quad (12)$$

$\lambda_1 = \lambda_0 = 1$ indique une protection totale de la vie privée. Cela signifie que la réponse donnée par l'unité répondante ne contient absolument aucune information sur le sujet étudié. Cela s'applique pour $a = 0$. Plus les mesures λ diffèrent de l'unité, plus la réponse figurant dans l'enregistrement contient d'information sur la caractéristique étudiée. Parallèlement, l'efficacité de l'estimation augmente (voir plus bas), mais la protection de l'individu contre l'enquêteur diminue. Dans le cas du plan d'interrogation directe avec $p_1 = 1$, où aucun masquage de la variable n'a lieu, ces mesures sont données par $\lambda_1 = \lambda_0 = \infty$.

Soit $\lambda_{1, opt}$ et $\lambda_{0, opt}$ les valeurs λ maximales de (11) et (12) qui, selon l'organisme statistique, permettent d'obtenir une protection suffisante des enregistrements contre la divulgation. En cas d'utilisation d'une stratégie en vue d'éviter la non-réponse et la réponse mensongère dans les enquêtes à collaborer sous forme d'une fonction de la protection perçue de la vie privée. Si la vie privée des

$$\hat{p}(\pi_A) = \frac{\pi_A \cdot (1 - \pi_A)}{N - n} \cdot \frac{n - 1}{N} \cdot \left(\frac{a^2}{b \cdot (1 - b)} + \frac{a}{1 - 2 \cdot b - a} \right) \cdot \pi_A \quad (9)$$

Afin de pouvoir calculer $\hat{\pi}_A$, la question sur l'appartenance à U_A (ou à U_A^c mais nous ignorons cette possibilité subéquemment sans perte de généralité) doit être incluse dans le plan d'interrogation à réponse aléatoire avec la probabilité $p_1 > 0$. Il existe, en tout, 16 combinaisons de cette question avec les quatre autres questions ou réponses (voir le tableau 1). Ces combinaisons peuvent être décrites comme des cas particuliers de notre stratégie standardisée de réponse. Par exemple, choisir $p_1 = 1$ mène à l'interrogation directe sur le sujet. Si nous posons que $0 < p_1 < 1$ et $p_2 = 1 - p_1$, le plan d'interrogation standardisé correspond à la procédure de Warner. Pour $0 < p_1 < 1$ et $p_3 = 1 - p_1$, nous obtenons la méthode d'Horvitz et ses collaborateurs avec la probabilité π_B connue (voir Greenberg, Abul-Elas, Simmons et Horvitz 1969). (Pour d'autres cas particuliers, déjà publiés autant que nous sachieons, nous renvoyons le lecteur à la colonne « Références » du tableau 1).

Tableau 1
Tous les cas particuliers de la stratégie standardisée fondée sur la réponse aléatoire

Plan	Questions/réponses	U _A	U _A ^c	U _B	Oui	Non	Références
S71	•	•	•	•	•	•	Interrogation directe
S72	•	•	•	•	•	•	Warner (1965) ¹
S73	•	•	•	•	•	•	Greenberg et coll. (1969) ²
S74	•	•	•	•	•	•	
S75	•	•	•	•	•	•	
S76	•	•	•	•	•	•	
S77	•	•	•	•	•	•	Quatember (2007) ³
S78	•	•	•	•	•	•	
S79	•	•	•	•	•	•	
S710	•	•	•	•	•	•	Singh, Hom, Singh et Mangat (2003) ⁴
S711	•	•	•	•	•	•	Fidler et Kleinmunch (1977) ⁵
S712	•	•	•	•	•	•	
S713	•	•	•	•	•	•	
S714	•	•	•	•	•	•	
S715	•	•	•	•	•	•	
S716	•	•	•	•	•	•	

1. Une version à deux degrés a été présentée par Mangat et Singh (1990).
2. Une version à deux degrés a été présentée par Mangat (1992).
3. Il s'agit d'une version à un degré de Mangat, Singh et Singh (1993).
4. Il s'agit d'une version à un degré de Singh, Mangat et Tracy (1994).
5. Une version à deux degrés a été présentée par Singh, Mangat et Mangat et Tracy (1995).

La question que soulèvent directement ces considérations est celle de savoir comment choisir les paramètres de plan de la méthode standardisée de réponse afin de couvrir les

de microdonnées doit être informé des détails de la méthode

de masquage.

À la section 2 du présent article, nous présentons une

nouvelle standardisation des méthodes de réponse aléatoire.

En outre, nous établissons les propriétés statistiques de l'esti-

mateur standardisé sous échantillonnage probabiliste général.

À la section 3, nous exposons la perspective essentielle de la

protection de la vie privée. À la section 4, nous répondons à

la question de savoir lequel des cas particuliers inclus dans la

standardisation est le plus efficace. À la section 5, nous

donnons un exemple fondé sur les données réelles, qui

illustre l'application des recommandations de la section 4

dans le contexte d'une enquête sur le comportement de

tricherie des étudiants.

2. Standardisation des stratégies fondées sur la réponse aléatoire

Soit la standardisation suivante des stratégies de randomi-
sation des réponses : chaque enquête doit répondre aléatoire-

ment avec la probabilité

- p_1 à la question « Êtes-vous membre du groupe U_1^A ? »
- p_2 à la question « Êtes-vous membre du groupe U_2^A ? » ou
- p_3 à la question « Êtes-vous membre du groupe U_B^A »

ou reçoit l'instruction de dire simplement

- « oui » avec la probabilité p_4 ou
- « non » avec la probabilité p_5

($\sum_{i=1}^5 p_i = 1$, $0 \leq p_i \leq 1$ pour $i = 1, 2, \dots, 5$). Les N_B éléments du groupe U_B sont caractérisés par la possession d'un attribut entièrement inoffensif B (par exemple, la saison B de naissance), qui ne devrait pas être relié à la possession ou à la non-possession de l'attribut A . Cette question non sensible sur l'appartenance au groupe U_B a été introduite comme une alternative à la question sur l'appartenance au groupe U_A par Horvitz, Shah et Simmons (1967) afin de réduire encore davantage la perception du caractère sensible de la procédure. $\pi_B = N_B/N$ (avec $0 < \pi_B < 1$) est la taille relative du groupe U_B . π_B et les probabilités p_1, p_2, \dots, p_5 sont les *paramètres de plan* de notre méthode standardisée de randomisation des réponses.

Soit

$$y_i = \begin{cases} 1 & \text{si l'unité } i \text{ répond « oui »,} \\ 0 & \text{autrement} \end{cases}$$

($i = 1, 2, \dots, n$). Pour un élément i la probabilité d'une réponse « oui » sous le plan d'interrogation à réponse aléatoire R est, sachant x :

$$P^R(y_i = 1) = p_1 \cdot x_1 + p_2 \cdot (1 - x_1) + p_3 \cdot \pi_B + p_4 \cdot \pi_B + p_5 \cdot \pi_B. \quad (4)$$

avec $a \equiv p_1 - p_2$ et $b \equiv p_2 + p_3 \cdot \pi_B + p_4 \cdot \pi_B + p_5 \cdot \pi_B$. Alors, le terme

$$x_1 = \frac{a}{y_i - b}$$

est sans biais pour la valeur réelle x_1 ($a \neq 0$). Si nous utilisons ces « substitués » pour x_1 (et en émettant l'hypothèse que la coopération des enquêtés est complète), les théorèmes qui suivent s'appliquent :

Théorème 1 : Sous un plan d'échantillonnage probabiliste avec probabilités d'inclusion π_i , nous avons l'estimateur

$$\hat{\pi}_A = \frac{1}{N} \cdot \sum_{i=1}^N \frac{x_i}{\pi_i}. \quad (5)$$

sans biais du paramètre π_A suivant :

Théorème 2 : Sous un plan d'échantillonnage probabiliste P , la variance de l'estimateur standardisé $\hat{\pi}_A$ (5) est donnée par

$$V^P(\hat{\pi}_A) = \frac{1}{N} \cdot \left(\sum_{i=1}^N \frac{x_i^2}{\pi_i} \right) + \left(\sum_{i=1}^N \frac{x_i^2}{\pi_i} \right) + \frac{a^2}{b \cdot (1 - b)} + \frac{a^2}{1 - 2 \cdot b - a} \cdot \sum_{i=1}^N \frac{x_i}{\pi_i} \cdot \pi_i. \quad (6)$$

Les preuves de ces deux théorèmes figurent à l'annexe. Le premier terme de la somme comprise entre les parenthèses extérieures de (6) fait référence à la variance de l'estimateur d'Horvitz-Thompson pour le total $\sum_{i=1}^N x_i$ sous un plan d'échantillonnage probabiliste P quand la question sur l'appartenance au groupe U_A est posée directement. Le deuxième terme de la somme peut être considéré comme le coût de la coopération des enquêtés en perte d'exactitude pour la protection de la vie privée offerte par le plan d'interrogation à réponse aléatoire. Apparemment, cet estimateur sans biais estimé sans biais en insérant l'estimateur sans biais $\hat{\pi}_A$ pour $V^P(\sum_{i=1}^N x_i / \pi_i)$ et $\sum_{i=1}^N x_i / \pi_i^2$ pour $\sum_{i=1}^N x_i^2 / \pi_i$.

Sous échantillonnage aléatoire simple sans remise, par exemple, l'estimateur (5) est donné par

$$\hat{\pi}_A = \frac{a}{\hat{\pi}_y - b} \quad (7)$$

avec $\hat{\pi}_y = \sum_{i=1}^N y_i / n$, la proportion de réponses « oui » dans l'échantillon. Dans ce cas, la variance (6) de l'estimateur standardisé $\hat{\pi}_A$ est donnée par

$$V(\hat{\pi}_A) = \frac{a^2}{\pi_A \cdot (1 - \pi_A)} \cdot \frac{N - 1}{N} + \frac{a^2}{b \cdot (1 - b)} + \frac{a^2}{1 - 2 \cdot b - a} \cdot \pi_A. \quad (8)$$

de ces derniers quant à la « confidentialité des données » et à la « protection perçue de la vie privée ». La première expression fait référence au désir qu'ont les enquêtés de voir leurs réponses demeurer hors de portée des personnes non concernées, tandis que la deuxième fait référence à leur souhait d'empêcher absolument tout le monde d'avoir accès à l'information. Singer, Mathiowetz et Cooper (1993), ainsi que Singer, van Hoewyk et Neugebauer (2003) signalent, à l'occasion de deux enquêtes successives auprès de la population américaine, que plus ces préoccupations sont vives, plus la probabilité de participer à l'enquête est faible (page 470ff et page 375ff).

Que peuvent apporter les statisticiens à ce domaine de recherche important ? Dans le cas de questions sensibles, l'utilisation de *stratégies fondées sur la réponse aléatoire* à l'étape de la conception de l'enquête peut réduire les taux de non-réponses et de réponses mensongères parce qu'elles donnent l'impression d'un accablissement de la protection des renseignements personnels. Une caractéristique commune de ces méthodes est que les questions directes sur le sujet sensible sont remplacées par un questionnaire conçu de telle manière que l'enquêteur n'est pas capable d'identifier la question (sélectionnée aléatoirement) à laquelle l'enquête a répondu, tout en permettant encore d'estimer le paramètre étudié. L'idée est de réduire de cette façon chez les enquêtés la crainte d'une « révélation » embarrassante et de s'assurer ainsi qu'ils seront disposés à coopérer. Pour atteindre cet objectif, l'enquête doit comprendre clairement comment la conception du questionnaire protège sa vie privée (voir Landsheer, van der Heijden et van Gils 1999, page 6ff).

Warner (1965). Dans son questionnaire, chaque personne interrogée devait répondre aléatoirement avec la probabilité p_1 à la question « Êtes-vous membre du groupe U_1 ? » ou avec la probabilité $p_2 = 1 - p_1$ à la question alternative « Êtes-vous membre du groupe U_2 ? » ($0 < p_1 < 1$). Depuis, différentes méthodes de réponse aléatoire fondées sur divers procédés de randomisation ont été proposées (pour une revue, consulter Chaudhuri et Mukerjee 1987, Nathan 1988, ou Tracy et Mangat 1996). Toutes ces stratégies s'appuient sur des questions ou des réponses sélectionnées aléatoirement, quoique certaines utilisent des procédés de randomisation différents selon que l'enquête possède ou non un attribut particulier (voir, par exemple, Kuk 1990 ; Mangat 1994 ; Kim et Warde 2005).

Warner (1971) a été le premier à constater que ces méthodes pouvaient aussi s'appliquer pour masquer des ensembles de microdonnées confidentielles afin de permettre leur grande diffusion (voir, ibidem, page 887). Ces ensembles de microdonnées peuvent contenir des variables donnant lieu à l'identification directe des unités étudiées, comme le nom ou un numéro d'identification, mais aussi

des variables fournissant des renseignements délicats sur une personne. Afin de protéger les unités étudiées contre la divulgation, il pourrait ne pas suffire de supprimer les variables auxquelles elles sont directement liées, parce que certaines unités pourraient encore être identifiées d'après le reste de leurs enregistrements. Le contrôle de la divulgation statistique n'est rien d'autre qu'un exercice d'équilibre entre la protection de l'anonymat des sujets participant à l'enquête et la préservation de l'information contenue dans les données (voir Skinner, Marsh, Openshaw et Wymer 1994). Les méthodes de masquage des données peuvent être réparties en trois catégories (voir Domingo-Ferrer et Mateo-Sanz 2002 ou Winkler 2004), à savoir 1) le *recodage global* des variables en des catégories moins détaillées ou de plus grands intervalles (voir par exemple, Willenborg et de Waal 1996, page 5f) ou le *recodage local* en utilisant divers scénarios de groupement au niveau de l'unité (voir Hua et Pei 2008, page 215f), 2) la *suppression locale* de certaines variables pour les unités étudiées présentant un risque élevé de réidentification en fixant simplement leur valeur à « manquante » (voir Willenborg et de Waal 1996, page 77) et 3) la *substitution* d'autres valeurs aux valeurs réelles d'une variables.

L'une des stratégies de la troisième catégorie est la *micro-agrégation* des variables (voir Defays et Anwar 1998). Dans ce cas, les vraies valeurs des variables sont, par exemple, triées par taille, puis réparties en (petits) groupes. Pour chaque groupe, des données agrégées sont diffusées au lieu des observations originales. Une autre méthode de ce type est la *permutation des données*, où celles provenant d'unités présentant un risque élevé de réidentification sont interchangeables avec des données provenant d'un autre ensemble d'unités étudiées (voir Dalenius et Reiss 1982). Une autre technique de substitution d'information identifi-catoire ou sensible est l'*ajout d'un bruit* aux valeurs observées, autrement dit l'ajout du résultat d'une expérience aléatoire à chaque donnée (voir Dalenius 1977 ou Fuller 1993). Enfin, les méthodes de randomisation des réponses peuvent aussi être utilisées pour masquer des variables identifiantes ou délicates. Dans ce cas, soit le masquage des données fournies par les unités échantillonnées est déjà effectué à l'étape de la conception de l'enquête, soit l'organisme statistique applique le mécanisme probabiliste de la méthode avant la diffusion du fichier de micro-données (voir Rosenberg 1980, Kim 1987, Gouweleuw, Kooiman, Willenborg et de Wolf 1998, ou van den Hout et van der Heijden 2002).

Toutes les méthodes de contrôle de la divulgation statistique protègent la vie privée des unités étudiées par une perte d'information qui peut être considérée comme le prix à payer pour cette protection. Afin de pouvoir corriger comme il convient le processus d'estimation, l'utilisateur du fichier

Une standardisation des stratégies fondées sur la réponse aléatoire

Andreas Quatember¹

Résumé

Les stratégies fondées sur la réponse aléatoire, qui ont été élaborées au départ à titre de méthodes statistiques destinées à réduire la non-réponse ainsi que la réponse mensongère, peuvent aussi être appliquées dans le domaine du contrôle de la divulgation statistique dans les fichiers de microdonnées à grande diffusion. Le présent article décrit une standardisation des méthodes de réponse aléatoire en vue d'estimer des proportions pour des attributs identificateurs ou sensibles. Les propriétés statistiques de l'estimateur standardisé sont établies dans le cas de l'échantillonnage probabiliste général. Afin d'analyser l'effet du choix des « paramètres de plan » implicites de la méthode sur la performance de l'estimateur, nous incluons dans l'étude des mesures de la protection de la vie privée. Pour cela, les variables réelles doivent être classées dans diverses catégories de sensibilité. Un exemple fondé sur des données réelles illustre l'application de la méthode à une enquête sur la tricherie chez les étudiants.

Mots clés : Protection de la vie privée ; contrôle de la divulgation statistique ; non-réponse ; réponse mensongère.

1. Introduction

Les cas de refus de répondre ou de donner la vraie réponse sont naturels dans les enquêtes par sondage. Ils peuvent donner lieu à un estimateur des paramètres de population présentant un biais de grandeur inconnue et une forte variance. Par conséquent, un utilisateur sérieux des données ne peut pas ignorer l'existence de la non-réponse et de la réponse mensongère.

Soit U l'univers de N unités de population et U_j un sous-ensemble de N_j éléments, qui appartiennent à une classe A d'une variable catégorique étudiée. En outre, soit U_j^c le groupe de N_j^c éléments qui n'appartiennent pas à cette classe ($U = U_j \cup U_j^c$, $U_j \cap U_j^c = \emptyset$, $N = N_j + N_j^c$). Soit

$$x_j = \begin{cases} 1 & \text{si l'unité } i \in U_j \\ 0 & \text{autrement} \end{cases}$$

($i = 1, 2, \dots, N$) et le paramètre d'intérêt π_j , qui est la taille relative de la sous-population U_j :

$$\pi_j = \frac{N_j}{N} = \frac{\sum_{i \in U_j} 1}{N} \quad (1)$$

($\sum_{i \in U_j} x_j$ est la notation abrégée de $\sum_{i \in U_j} 1$). Dans le cas d'un échantillon probabiliste s (voir par exemple, Sæmstad, Swensson et Wretman 1992, page 81), un estimateur de π_j peut être calculé à partir de l'estimateur d'Horvitz-Thompson de N_j par

$$\hat{\pi}_j^{\text{HT}} = \frac{1}{n} \cdot \sum_{i \in s} x_j \pi_i \quad (2)$$

($\pi_i > 0$ est la probabilité que l'unité i soit incluse dans l'échantillon), si la question « Êtes-vous un membre du groupe U_j ? » (ou une question équivalente) est posée directement (dir). Cet estimateur est sans biais si toutes les observations x_j ($i = 1, 2, \dots, n$) sont des réponses sincères. En présence de non-réponse totale ou partielle en ce qui concerne une variable étudiée, l'échantillon s est divisé en un « ensemble de réponses » $r \subset s$ de taille n_r et un « ensemble de réponses manquantes » $m \subset s$ de taille n_m ($s = r \cup m$, $r \cap m = \emptyset$, $n = n_r + n_m$). Dans le cas de variables d'un caractère hautement personnel, embarrassant comme la toxicomanie, les maladies, le comportement sexuel, la fraude fiscale, l'alcoolisme, la violence familiale ou la criminalité), r est en outre divisé en un ensemble t de n_t unités échantillonnées qui répondent sincèrement, et un ensemble u de taille n_u , d'unités qui répondent de manière mensongère ($r = t \cup u$, $t \cap u = \emptyset$, $n_r = n_t + n_u$). L'estimateur (2) doit alors être réécrit sous la forme :

$$\hat{\pi}_j^{\text{HT}} = \frac{1}{n} \cdot \left(\sum_{i \in t} \frac{\pi_i}{x_i} + \sum_{i \in u} \frac{\pi_i}{x_i} + \sum_{i \in m} \frac{\pi_i}{x_i} \right) \quad (3)$$

Naturellement, les éléments de l'ensemble u ne peuvent pas être identifiés et les x_i de m ne sont pas observables, ce qui introduit des erreurs de mesure et de non-réponse dans l'estimation. Par conséquent, tout doit être fait en vue de maintenir les taux de réponses mensongères et de non-réponses aussi faibles que possible.

Les caractéristiques du plan de sondage, qui ont manifestement une incidence sur la qualité de l'information demandée aux enquêtés (voir par exemple Groves, Fowler, Lepkowski, Singer et Tourangeau 2004, Section 6.7), sont étroitement liées aux préoccupations

- Skirken, M.G. (2004). Network sample surveys of rare and elusive populations: A historical review. Dans *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Disponible au <http://www.statcan.gc.ca/pub/11-522-x/2004001/8614-4-eng.pdf>.
- Skirken, M.G. (2005). Network sampling developments in survey research during the past 40+ years. *Survey Research*, 36, 1, 1-5. Disponible au <http://www.srli.unc.edu/Publist/Newsletter/pastissues.htm>.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Smith, P.J., Battaglia, M.P., Huggins, V.J., Hoaglin, D.C., Roden, A., Khare, M., Ezzati-Rice, T.M., et Wright, R.A. (2001). Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine*, 20(4S), 17-24.
- Statistique Canada (2008). *Enquête sur la santé dans les collectivités canadiennes (E5SC)*. Disponible au <http://www.statcan.gc.ca/cgi-bin/mmbp/DSV.f?function=getSurvey&SDDS=3226&lang=en&db=mmbp&adm=8&dis=2>.
- Statman, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association*, 67, 335-339.
- Statman, S. (1976). *Applied Sampling*. New York : Academic Press.
- Statman, S., et Freeman, H.E. (1988). The use of network sampling for locating the seriously ill. *Medical Care*, 26, 992-999.
- Statman, S., et Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.
- Statman, S., Sirken, M.G., et Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.
- Thompson, S.K. (2002). *Sampling*. 2^{ème} Edition. New York : John Wiley & Sons, Inc.
- Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépiégeage de liens. *Techniques d'enquête*, 26, 99-112.
- Thompson, S.K., et Seber, G.A.F. (1996). *Adaptive Sampling*. New York : John Wiley & Sons, Inc.
- Tortora, R., Groves, R.M., et Peytcheva, E. (2008). Multiplicity-based sampling for the mobile telephone population: Coverage, nonresponse, and measurement issues. Dans *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japel, P.J. Lavrakas, M.W. Link et R.L. Sangster). Hoboken, NJ : Wiley, 133-148.
- U.S. Census Bureau (2009a). *Design and Methodology, American Community Survey*. U.S. Government Printing Office, Washington, DC.
- Skirken, M.G. (2005). Network sampling developments in survey research during the past 40+ years. *Survey Research*, 36, 1, 1-5. Disponible au <http://www.srli.unc.edu/Publist/Newsletter/pastissues.htm>.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Smith, P.J., Battaglia, M.P., Huggins, V.J., Hoaglin, D.C., Roden, A., Khare, M., Ezzati-Rice, T.M., et Wright, R.A. (2001). Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine*, 20(4S), 17-24.
- Statistique Canada (2008). *Enquête sur la santé dans les collectivités canadiennes (E5SC)*. Disponible au <http://www.statcan.gc.ca/cgi-bin/mmbp/DSV.f?function=getSurvey&SDDS=3226&lang=en&db=mmbp&adm=8&dis=2>.
- Statman, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association*, 67, 335-339.
- Statman, S. (1976). *Applied Sampling*. New York : Academic Press.
- Statman, S., et Freeman, H.E. (1988). The use of network sampling for locating the seriously ill. *Medical Care*, 26, 992-999.
- Statman, S., et Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.
- Statman, S., Sirken, M.G., et Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.
- Thompson, S.K. (2002). *Sampling*. 2^{ème} Edition. New York : John Wiley & Sons, Inc.
- Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépiégeage de liens. *Techniques d'enquête*, 26, 99-112.
- Thompson, S.K., et Seber, G.A.F. (1996). *Adaptive Sampling*. New York : John Wiley & Sons, Inc.
- Tortora, R., Groves, R.M., et Peytcheva, E. (2008). Multiplicity-based sampling for the mobile telephone population: Coverage, nonresponse, and measurement issues. Dans *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japel, P.J. Lavrakas, M.W. Link et R.L. Sangster). Hoboken, NJ : Wiley, 133-148.
- U.S. Census Bureau (2009a). *Design and Methodology, American Community Survey*. U.S. Government Printing Office, Washington, DC.
- U.S. Census Bureau (2009b). *Small Area Income and Poverty Estimates*. Disponible au <http://www.census.gov/did/www/saipc/methods/statecounty/index.html>.
- U.S. National Center for Health Statistics (2009a). *National Health Interview Survey (NHIS)*. Disponible au <http://www.cdc.gov/nchs/nhis/methods.htm>.
- U.S. National Center for Health Statistics (2009b). *The National Immunization Survey (NIS)*. Disponible au <http://www.cdc.gov/nis/about-eng.htm>.
- U.S. National Center for Health Statistics (2009c). *State and Local Area Integrated Telephone Survey (SLAITS)*. Disponible au <http://www.cdc.gov/nchs/about/major/stats/nslch.htm>.
- Volz, E., et Heckathorn, D.J. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24, 79-97.
- Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 429-434.
- Waksberg, J., Brick, J.M., Shapiro, G., Flores Cervantes, I. et Bell, B. (1997). Dual-frame RPD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-718.
- Waksberg, J., Judkins, D. et Massey, J.T. (1997). Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis. *Techniques d'enquête*, 23, 69-80.
- Watters, J.K., et Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.
- Wingiele, M., Park, I., Rusz, K., Liu, B. et Shapiro, G. (2007). A case study in dual-frame estimation methods. *American Statistical Association*, 3195-3202.
- Word, D.T., et Perkins, R.C. (1996). *Building a Spanish Surname List for the 1990's - A New Approach to an Old Problem*. Document de travail technique de la Division de la population, No. 13. U.S. Census Bureau, Washington, DC.
- Wu, C., et Rao, J.N.K. (2009). Empirical likelihood methods for inference from multiple frame surveys. *Proceedings of the International Statistical Institute, Durban, South Africa*.
- Xia, Q., Tholandi, M., Osmond, D.H., Pollack, L.M., Zhou, W., Ruiz, J.D. et Catania, J.A. (2006). The effect of venue sampling on estimates of HIV prevalence and sexual risk behaviors in men who have sex with men. *Sexually Transmitted Diseases*, 33, 545-550.

- Kanouse, D.E., Berry, S.H. et Duan, N. (1999). Drawing a probability sample of female street prostitutes in Los Angeles County. *Journal of Sex Research*, 36, 45-51.
- Katzoff, M.J. (2004). Applications of adaptive sampling procedures to problems in public health. Dans *Proceedings of Statistics Canada Symposium 2004, Innovative Methods for Surveying Difficult-to-Reach Populations*. Disponible au <http://www.statcan.gc.ca/pub/11-522-x/2004001/8751-eng.pdf>
- Katzoff, M.J., Sirken, M.G. et Thompson, S.K. (2002). Proposals for adaptive and link-tracing sampling designs in health surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1772-1775.
- Kish, L. (1965a). Selection techniques for rare traits. Dans *Genetics and the Epidemiology of Chronic Diseases*. Public Health Service Publication No. 1163.
- Kish, L. (1965b). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- Kish, L. (1976). Optima and proxima in linear sample design. *Journal of the Royal Statistical Society, A*, 139, 80-95.
- Kish, L. (1987). *Statistical Design for Research*. New York : John Wiley & Sons, Inc.
- Kish, L. (1988). Plans de sondage à usages multiples. *Techniques d'enquête*, 14, 19-33.
- Kish, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques. *Techniques d'enquête*, 25, 147-158.
- Körner, T., et Nimmergut, A. (2004). A permanent sample as a sampling frame for difficult-to-reach populations? Dans *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Disponible au <http://www.statcan.gc.ca/pub/11-522-x/2004001/8752-eng.pdf>
- Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heerenga, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fulz, N.H., Hund, M.D., Potter, G.G., Rodgers, W.L., Stephens, D.C., Weir, D.R. et Willis, R.J. (2005). The Aging Demographics, and Memory Study: Study design and methods. *Neuroepidemiology*, 25, 181-191.
- Lauderdale, D.S. et Keiselman, B. (2000). Asian American ethnic identification by surname. *Population Research and Policy Review*, 19, 283-300.
- Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.
- Lavallée, P. (2007). *Indirect Sampling*. New York : Springer.
- Lohr, S.T. (2009). Multiple-frame surveys. Dans *Handbook of Statistics, Volume 29A: Sample Surveys: Design, Methods, and Applications*. (Eds., D. Pfeffermann et C.R. Rao). Burlington, MA : Elsevier B.V.
- Lohr, S.T., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.T., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, 97-106.
- Mackellar, D., Vallero, L., Karon, J., Lemp, G. et Janssen, R. (1996). The Young Men's Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111, Supplément 1, 138-144.
- Maffeo, C., Frey, W. et Kaiton, G. (2000). Survey design and data collection in the disability evaluation study. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 79-88.
- Marker, D.A. (2001). Production d'estimations régionales d'après les données d'enquêtes nationales : méthodes visant à réduire au minimum l'emploi d'estimateurs indirects. *Techniques d'enquête*, 27, 201-207.
- McKenzie, D.J., et Misiäinen, J. (2009). Surveying migrant households: A comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society*, 172, 339-360.
- Mecca, F. (2004). Center sampling: A strategy for sampling difficult-to-sample populations. Dans *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Disponible au <http://www.statcan.gc.ca/pub/11-522-x/2004001/8740-eng.pdf>
- Metcalf, P., et Scott, A. (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28, 1512-1523.
- Michael, R.T., et O'Muirheartaigh, C.A. (2008). Design priorities and disciplinary perspectives: The case of the US National Children's Study. *Journal of the Royal Statistical Society, A*, 171, 465-480.
- Mohadjer, L.R., et Curtin, L.R. (2008). Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey. *Techniques d'enquête*, 34, 119-126.
- Morris, P. (1965). *Prisoners and Their Families*. Allen and Unwin, Londres, 303-306.
- National Children's Study (2007). Study design. Dans *The National Children's Study Research Plan*. Version 1.3. Disponible au <http://www.nationalchildrensstudy.gov/research/studydesign/resca> <http://www.nationalchildrensstudy.gov/research/studydesign/resca>
- Pew Research Center (2007). *Muslim Americans, Middle Class and Mostly Mainstream*. Pew Research Center, Washington, DC.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Rodriguez Vera, A. (1982). Multipurpose optimal sample allocation using mathematical programming. Biostatistics Doctoral Dissertation. Ann Arbor : University of Michigan.
- Roehbart, G.S., Fine, M. et Sudman, S. (1982). On finding and interviewing the needles in the haystack: The use of multiplicity sampling. *Public Opinion Quarterly*, 46, 408-421.
- Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : problèmes et solutions. *Techniques d'enquête*, 20, 3-23.
- Statistique Canada, N° 12-001-X au catalogue

- tior, C.F., et Kalton, G. (Eds.) (2007). *Using the American Community Survey: Benefits and Challenges*. Washington, DC: National Academies Press.
 Clark, R.G., et Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society*, 170, S enes A, 63-82.
 Defrances, C.J., Lucas, C.A., Butte, V.C., et Golosinskiy, A. (2008). 2006 National Hospital Discharge Survey. National Health Statistics Reports Number 5. U.S. National Center for Health Statistics, Hyattsville, MD.
 Deming, W.E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *Revue Internationale de Statistique*, 45, 29-37.
 Durr, J.-M. (2005). The French new rolling census. *Statistical Journal of the United Nations Economic Commission for Europe*, 22, 3-12.
 Elliott, M.N., Finch, B.K., Klein, D., Ma, S., Do, D.P., Becker, M.K., Orr, N., et Lurie, N. (2008). Sample designs for measuring the health of small racial/ethnic subgroups. *Statistics in Medicine*, 27, 4016-4029.
 Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantaja, P., et Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services Research Methods*, 9, 69-83.
 Elliott, M.N., McCaffrey, D., Perlman, J., Marshall, G.N., et Hambergsmoomian, K. (2009). Use of expert ratings as sampling strata for a more cost-effective probability sample of a rare population. *Public Opinion Quarterly*, 73, 56-73.
 Erens, B., Prior, G., Koroveriss, C., Calderwood, L., Brookes, M., et Primatesta, P. (2001). Survey methodology and response. In *Health Survey for England - The Health of Minority Ethnic Groups*, 99. Volume 2: Methodology and Documentation. (Eds. B. Erens, P. Primatesta et G. Prior). The Stationery Office, Londres.
 Fecso, R.S., Baskin, R., Chu, A., Gray, C., Kalton, G., et Phelps, R. (2007). *Design Options for SESTAT for the Current Decade*. Document travail SRS 07-021. Division of Science Resource Statistics, U.S. National Science Foundation.
 Fiscella, K., et Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41, 1482-1500.
 Flores Cervantes, I., et Kalton, G. (2008). Methods for sampling rare populations in telephone surveys. Dans *Advances in Telephone Survey Methodology*. (Eds. J.M. Brick, E.D. de Leeuw, P. Lape , P. Lavrakas, M.W. Link et R.L. Sangster). Hoboken, NJ : Wiley, 113-132.
 Folsom, R.E., Potter, F.J., et Williams, S.K. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 792-796.
 German Federal Statistical Office (2009). *Microcensus*. Disponible au http://www.destatis.de/jsp/desktop.do?cid=desstat1/inner/en/EN/press/abstsz/Mikrocensus_c...&templateId=renderPrint.pml.
 Kalton, G., et Brick, J.M. (1995). M ethodes de pond eration pour les enqu tes par panel aupr s des m nages. *Techniques d'enqu te*, 21, 37-49.
 Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, 4, 149, 65-82.
 Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition*, 6, 491-501.
 Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. Dans *Methodological Issues in AIDS Behavioral Research*. (Eds. D.G. Ostrow, R.C. Kessler). New York : Plenum Press.
 Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, United Nations, New York.
 Kalton, G. (1991). L' chantillonnage des flux de populations humaines mobiles. *Techniques d'enqu te*, 17, 197-210.
 Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
 Kalton, G. (1991). L' chantillonnage des flux de populations humaines mobiles. *Techniques d'enqu te*, 17, 197-210.
 Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
 Longman, 245-261.
 Hedges, B.M., (1979). Sampling minority populations. Dans *Social and Educational Research in Action* (Ed., M.J. Wilson) Londres : Longman, 245-261.
 Hedges, B.M. (1973). *Sampling Minority Groups*. Thomson Medal Awards. Thomson Organization, Londres.
 Heckathorn, D.D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37, 151-208.
 Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
 Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhy *, 36, 99-118.
 Haerem, A.F., Anderson, D.W., et Schoenberg, B.S. (1986). Prevalence and clinical features of epilepsy in a biracial United States population. *Epilepsia*, 27, 66-75.
 Green, J. (2000). Mathematical programming for sample design and allocation problems. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 688-692.
 Gonzalez, J.F., Ezziati, T.M., White, A.A., Massey, J.T., Lago, J., et Waksberg, J. (1985). Sample design and estimation procedures. Dans *Plan and Operation of the Hispanic Health and Nutrition Examination Survey, 1982-84*. (Ed. K.R. Maurer). Vial and Health Statistics, Series I, No. 19. U.S. Government Printing Office, Washington, DC, 23-32.
 Goldman, J.D., Bernal, L.G., et Berlin, M. (1997). An overview of the USDA's 1994-96 Continuing Survey of Food intakes by individuals and the Diet and Health Knowledge Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 796-801.
 Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
 Heckathorn, D.D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37, 151-208.
 Hedges, B.M. (1973). *Sampling Minority Groups*. Thomson Medal Awards. Thomson Organization, Londres.
 Hedges, B.M., (1979). Sampling minority populations. Dans *Social and Educational Research in Action* (Ed., M.J. Wilson) Londres : Longman, 245-261.
 Hornig, M., Moore, W., Pedlow, S., et Wolter, K. (1999). Undercoverage in a large national screening survey for youths. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 570-575.
 Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
 Kalton, G. (1991). L' chantillonnage des flux de populations humaines mobiles. *Techniques d'enqu te*, 17, 197-210.
 Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, United Nations, New York.
 Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. Dans *Methodological Issues in AIDS Behavioral Research*. (Eds. D.G. Ostrow, R.C. Kessler). New York : Plenum Press.
 Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition*, 6, 491-501.
 Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, 4, 149, 65-82.
 Kalton, G., et Brick, J.M. (1995). M ethodes de pond eration pour les enqu tes par panel aupr s des m nages. *Techniques d'enqu te*, 21, 37-49.
 German Federal Statistical Office (2009). *Microcensus*. Disponible au http://www.destatis.de/jsp/desktop.do?cid=desstat1/inner/en/EN/press/abstsz/Mikrocensus_c...&templateId=renderPrint.pml.
 Folsom, R.E., Potter, F.J., et Williams, S.K. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 792-796.
 Flores Cervantes, I., et Kalton, G. (2008). Methods for sampling rare populations in telephone surveys. Dans *Advances in Telephone Survey Methodology*. (Eds. J.M. Brick, E.D. de Leeuw, P. Lape , P. Lavrakas, M.W. Link et R.L. Sangster). Hoboken, NJ : Wiley, 113-132.
 Fiscella, K., et Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41, 1482-1500.
 Fecso, R.S., Baskin, R., Chu, A., Gray, C., Kalton, G., et Phelps, R. (2007). *Design Options for SESTAT for the Current Decade*. Document travail SRS 07-021. Division of Science Resource Statistics, U.S. National Science Foundation.
 Fiscella, K., et Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41, 1482-1500.
 Flores Cervantes, I., et Kalton, G. (2008). Methods for sampling rare populations in telephone surveys. Dans *Advances in Telephone Survey Methodology*. (Eds. J.M. Brick, E.D. de Leeuw, P. Lape , P. Lavrakas, M.W. Link et R.L. Sangster). Hoboken, NJ : Wiley, 113-132.
 Folsom, R.E., Potter, F.J., et Williams, S.K. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 792-796.
 German Federal Statistical Office (2009). *Microcensus*. Disponible au http://www.destatis.de/jsp/desktop.do?cid=desstat1/inner/en/EN/press/abstsz/Mikrocensus_c...&templateId=renderPrint.pml.
 Kalton, G., et Brick, J.M. (1995). M ethodes de pond eration pour les enqu tes par panel aupr s des m nages. *Techniques d'enqu te*, 21, 37-49.
 Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition*, 6, 491-501.
 Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. Dans *Methodological Issues in AIDS Behavioral Research*. (Eds. D.G. Ostrow, R.C. Kessler). New York : Plenum Press.
 Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, United Nations, New York.
 Kalton, G. (1991). L' chantillonnage des flux de populations humaines mobiles. *Techniques d'enqu te*, 17, 197-210.
 Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
 Kalton, G. (1991). L' chantillonnage des flux de populations humaines mobiles. *Techniques d'enqu te*, 17, 197-210.
 Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
 Longman, 245-261.
 Hedges, B.M., (1979). Sampling minority populations. Dans *Social and Educational Research in Action* (Ed., M.J. Wilson) Londres : Longman, 245-261.
 Hedges, B.M. (1973). *Sampling Minority Groups*. Thomson Medal Awards. Thomson Organization, Londres.

aujourd'hui, plusieurs panels Internet et probabilités qui peuvent remplir ce rôle (Callagaro et DiSogra 2008). Toutefois, une réserve sérieuse au sujet de ce genre de panels tient aux faibles taux de réponse qui sont généralement obtenus.

4. Conclusion

Le présent article donne un bref aperçu de la gamme de méthodes utilisées dans les enquêtes par sondage pour échantillonner et suréchantillonner les populations rares, principalement celles classées par Kish comme étant des petits domaines (les références citées fournissent plus de renseignements). Bien que les méthodes soient examinées individuellement, en pratique, elles sont souvent combinées, surtout quand il existe plusieurs domaines rares d'intérêt.

réalisée par téléphone, on a combiné la stratification disproportionnée (suréchantillonnage des centraux téléphoniques pour lesquels la prévalence des populations coréennes et vietnamiennes d'intérêt est plus élevée) et un plan à base de sondage double (méthode de composition aléatoire (CA) complétée par une liste des noms probablement coréens et vietnamiens). Dans de nombreux cas, l'art de la création d'un plan d'échantillonnage probabiliste efficace pour une population rare réside dans l'application créative d'une combinaison de méthodes.

L'enquête téléphonique auprès des Américains musulmans réalisée par le Pew Research Center est un autre exemple, trois méthodes d'échantillonnage ayant été employées pour échantillonner cette population très rare (Pew Research Center 2007). L'une des composantes du plan de sondage était un échantillon obtenu par CA stratifié géographiquement, avec échantillonnage disproportionnel- lement stratifié dans les strates définies en fonction de la prévalence des Américains musulmans. La strate dans laquelle la prévalence était la plus faible a été traitée comme strate seuil et exclue. La deuxième composante était un échantillon de reprise de contact avec des Américains musulmans tiré de la base de données sur les interviews des enquêtes récentes du Pew Research Center. La troisième composante était un échantillon obtenu par CA à partir d'une liste d'Américains vraisemblablement musulmans produite par un fournisseur commercial. Pour éviter les pistes de sélection en double entre les strates géographiques et la liste du fournisseur commercial, les numéros de téléphone sélectionnés dans les strates géographiques ont été apparés à la liste du fournisseur commercial et supprimés de l'échantillon issus des strates géographiques pendant un appartement était découvert.

Non seulement les diverses techniques d'échantillonnage sont fréquemment combinées dans les plans d'échantillonnage pour populations rares, mais plusieurs de ces

Bibliographie

J'aimerais remercier Daniel Levine et Leyla Mohadjer pour leur revue utile d'une ébauche de cet article, Daifeng Han et Army Lin de leurs commentaires constructifs concernant une version antérieure, plus courte, ainsi que Mike Brick, Marc Elliott et Jon Rao de leurs conseils concernant des points particuliers.

Remerciements

techniques sont interdépendantes. Ainsi, les bases de sondage multiples peuvent être traitées par la méthode d'identification unique (voir la section 3.4), qui, en fait, est simplement une stratification disproportionnée. Tandis que la population complète est répartie en strates pour la stratification disproportionnée, la même approche est adoptée pour l'échantillonnage à deux phases, mais la classification dans les strates est appliquée uniquement aux membres de l'échantillon de première phase. La théorie de l'échantillonnage par réseaux est semblable à celle de l'échantillonnage à bases multiples, quand cette dernière méthode s'appuie sur les probabilités de sélection globales inverses comme pondération dans l'analyse. Ces interdépendances expliquent les similitudes observées dans les fondements théoriques de ces méthodes.

- Anderson, D.W., et Kellon, G. (1990). Case-finding strategies for studying rare chronic diseases. *Statistica Applicata*, 2, 309-321.
- Ardliff, P., et Le Blanc, D. (2001). Échantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français. *Techniques d'enquête*, 27, 117-127.
- Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *American Statistician*, 42, 174-177.
- Bolling, K., Grant, C. et Sinclair, P. (2008). 2006-07 British Crime Survey (England and Wales). Rapport technique. Volume 1. Disponible au <http://www.homeoffice.gov.uk/rds/pdfs07/bcs0607tech1.pdf>.
- Brick, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.

- Callegaro, M., et Disogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 1008-1032.
- Camburn, D. P., et Wright, R. A. (1996). *Predicting eligibility rates for rare populations in RDD screening surveys*. Disponible au <http://www.cdc.gov/nids/pdfs/design/camburn1996.pdf>.
- Chn, A., Brick, J. M., et Kalton, G. (1999). Weights for combining surveys across time or space. *Bulletin of the International Statistical Institute, Contributed Papers*, 2, 103-104.

Lorsque l'on envisage l'échantillonnage de populations rares dans le contexte d'enquêtes par panel, il importe de faire la distinction entre celles qui sont définies par des caractéristiques statiques et celles définies pour des caractéristiques non statiques. Aucune accumulation au cours du temps ne peut être effectuée dans le cas d'enquêtes par panel pour des populations rares définies en fonction de caractéristiques statiques, telles que la race ou l'ethnicité. Cependant, si un échantillon d'une population rare statique est tiré à un point particulier dans le temps, il peut être utile de suivre cet échantillon dans un panel pour étudier les caractéristiques de cette population à des points ultérieurs dans le temps, éventuellement en ajoutant des échantillons supplémentaires pour représenter les personnes qui sont entrées dans la population après la sélection de l'échantillon original. Fecso, Baskin, Chu, Gray, Kalton et Phelps (2007) décrivent comment cette approche a été appliquée à l'échantillonnage des scientifiques et des ingénieurs américains au cours d'une décennie. Pour les années 1990, la National Survey of College Graduates (NSCG) a été réalisée en 1993 auprès d'un échantillon stratifié de titulaires d'un diplôme collégial sélectionnés d'après les enregistrements de l'échantillon du Recensement de la population de 1990 ayant reçu le questionnaire complet. Les personnes qui se sont avérées être des scientifiques ou des ingénieurs ont ensuite été étudiées de nouveau dans le cadre des cycles de 1995, 1997 et 1999 de la NSCG. Pour représenter les nouveaux entrants dans la population cible, une autre enquête – la Survey of Recent College Graduates – a été réalisée les mêmes années que la NSCG. Un sous-échantillon des diplômés récents des collèges a été ajouté au cycle suivant du panel de la NSCG à chaque occasion.

Les enquêtes par panel peuvent être utilisées pour accumuler des échantillons de populations rares non statiques, principalement des personnes vivant des événements tels qu'une naissance ou un divorce. Par exemple, aux États-Unis, dans le cadre de la National Children's Study, il est prévu de suivre un grand échantillon de femmes en âge de procréation admissibles au cours d'une période d'environ quatre ans et de faire participer celles qui deviennent enceintes à l'étude principale, qui est une étude longitudinale Children's Study 2007, Michael et O'Muircheartaigh (2008). Enfin, un grand échantillon peut être recruté dans un panel et fournir des données qui permettent d'identifier les membres de diverses populations rares qui pourraient présenter un intérêt dans l'avenir. Ces membres sont alors suivis dans le panel et en fonction de leur appartenance à une population rare, inclus dans les échantillons des enquêtes pour lesquelles ils sont qualifiés. Körner et Nimmervegt (2004) décrivent un « panel d'accès » allemand qui pourrait être utilisé de cette façon, et il existe

fréquemment avaient des taux plus élevés de comportements sexuels à risque élevé que les autres et que les taux de résultats ont attiré l'attention sur le fait qu'il est difficile de générer un échantillon représentatif par échantillonnage des lieux de sociabilité.

McKenzie et Mistiaen (2009) ont procédé à une expérience en vue de comparer l'échantillonnage des lieux de sociabilité (échantillonnage par interception) à l'échantillonnage aréolaire ainsi qu'à l'échantillonnage en boule de neige pour l'échantillonnage des Brésiliens de descendance japonaise (Nikkei) à Sao Paulo et à Paraná. Les lieux de sociabilité comprennent des endroits où les Nikkei se rendent souvent (par exemple un club de sport, une station de métro, des épiceries et un club culturel japonais) et des événements (par exemple, un film japonais et un festival d'alimentation japonaise). Sur la base de cette expérience, ils ont conclu que l'échantillonnage des lieux de sociabilité (et l'échantillonnage en boule de neige) produisait un suréchantillonnage des personnes les plus étroitement liées à la communauté nikkei et ne donnait donc pas des échantillons représentatifs. Ce résultat non surprenant met en relief les réserves concernant l'utilisation de l'échantillonnage des lieux de sociabilité pour échantillonner des visites à des lieux particuliers.

3.7 Accumulation ou rétention d'échantillons au cours du temps

Si la collecte des données d'enquête se répète au cours du temps, les concepts d'enquêtes peuvent tirer parti de cette caractéristique pour l'échantillonnage des populations rares (Kish 1999). Il importe de souligner une distinction importante entre les enquêtes répétées et les enquêtes par panel. Des échantillons de membres d'une population rare peuvent être facilement accumulés au fil du temps dans le cas d'enquêtes répétées. Par exemple, aux États-Unis, la National Health Interview Survey est réalisée hebdomadairement auprès d'échantillons représentatifs de la population nationale; des échantillons de populations rares peuvent être accumulés sur une ou plusieurs années jusqu'à ce qu'une taille d'échantillon suffisante soit atteinte (U.S. National Center for Health Statistics 2009a). Dans le cas de l'accumulation au cours du temps, les estimations produites sont des estimations par période plutôt que ponctuelles qui peuvent être difficiles à interpréter quand les caractéristiques à analyser varient fortement au cours du temps (Citró et Kalton 2007). Ainsi, comment faut-il interpréter un taux de pauvreté sur une période de trois ans pour une population minoritaire rare quand le taux de pauvreté a varié fortement au cours de la période ?

parce que l'on n'a pas réussi à rejoindre ces jeunes, particulièrement ceux vivant dans un ménage n'ayant pas le téléphone.

Tortora, Groves et Peytcheva (2008) donnent un autre exemple d'utilisation de l'échantillonnage par réseau, cette fois-ci pour essayer de couvrir des personnes ne possédant qu'un téléphone mobile au moyen d'un échantillon obtenu par CA de numéros de téléphone conventionnel à fil. On a demandé aux répondants à l'enquête par CA (elle-même une enquête par panel) de fournir de renseignements sur leurs parents, leurs frères et sœurs et leurs enfants adultes vivant dans un ménage doté uniquement d'un téléphone mobile. Ces résultats illustrent certains problèmes généraux que pose l'échantillonnage par réseaux, savoir si les membres du réseau possédaient seulement un téléphone mobile dépendait de la cohésion du réseau, la réticence à communiquer les numéros de téléphone mobile était généralisée et nombre des ménages identifiés comme ne possédant qu'un téléphone mobile possédaient en fait également un téléphone conventionnel à fil.

Dans la pratique, l'échantillonnage par réseaux n'a pas été utilisé à grande échelle pour les enquêtes sur les membres des populations rares. Certaines limites de la méthode sont illustrées par les études décrites plus haut. Il existe un risque que la personne échantillonnée pour fournir l'information ne déclare pas correctement la situation d'appartenance à la population rare d'autres membres du réseau, délibérément ou par manque d'information. La non-réponse à l'étape de la collecte principale des données est une autre préoccupation. En outre, des questions d'ordre éthique peuvent se poser quand on demande aux personnes échantillonnées d'indiquer si les membres de leur réseau appartiennent à la population rare, quand cette appartenance est de nature sensible. Les avantages de l'échantillonnage par réseaux sont partiellement annulés par l'accroissement des erreurs d'échantillonnage dues à la pondération variable que comporte la méthode et par les coûts de repérage des membres liés de la population rare.

3.6 Échantillonnage des lieux de sociabilité

L'échantillonnage des lieux de sociabilité est utilisé à grande échelle afin d'échantillonner les populations qui n'ont pas de lieu de résidence fixe pour les recensements ainsi que les enquêtes. Les nomades mènent leurs animaux aux abreuvoirs et les sans-logis peuvent être échantillonnés dans les locaux des soupes populaires quand ils s'y rendent pour obtenir de la nourriture (par exemple Kalton 1993a; Ardlly et Le Blanc 2001). Une caractéristique essentielle de l'échantillonnage des lieux de sociabilité est qu'il comporte un élément temporel (période), ce qui crée des problèmes de multiplicité (Kaisbeek 2003). Une réserve importante concernant la méthode est qu'elle ne permet pas de couvrir

L'échantillonnage des lieux de sociabilité est souvent été utilisé pour échantillonner les hommes ayant des relations sexuelles avec des hommes, les lieux de sociabilité étant les lieux que fréquentent ces hommes, tels que les bars, bains publics et librairies fréquentés par les gays (Kalton 1993b, MacKellar, Valleroy, Karon, Lemp et Janssen 1996). En Tholandi, Osmond, Pollack, Zhou, Ruiz et Catania (2006) ont constaté que les hommes qui visitaient les lieux gays

les personnes qui ne visitent aucun des lieux spécifiés durant la période particulière choisie.

L'échantillonnage axé sur les lieux de sociabilité est utilisé pour échantillonner les populations rares mobiles, telles que les passagers des aéroports ou les visiteurs d'un musée ou d'un parc national. Dans de tels cas, la question qui se pose est celle de savoir si l'unité d'analyse est la visite ou le visiteur. Quand la visite est l'unité appropriée, il n'existe aucun problème de multiplicité (voir, par exemple, le rapport sur la National Hospital Discharge Survey réalisée aux États-Unis publié par DeFrances, Lucas, Buie et Golosinsky 2008). Par contre, si le visiteur est l'unité d'analyse, le fait qu'un visiteur peut faire plusieurs visites durant la période choisie doit être pris en compte (Kalton 1991; Sudman et Kalton 1986). Une approche consiste à traiter les visites comme étant admissibles uniquement s'il s'agit des premières visites faites durant la période de référence de l'enquête. Une autre approche consiste à appliquer aux pondérations des corrections pour tenir compte de la multiplicité dans l'analyse; cependant, déterminer le nombre de visites effectuées pose des problèmes, parce que certaines auront lieu après celle qui est échantillonnée.

L'échantillonnage des lieux de sociabilité a également été utilisé pour échantillonner une gamme de populations rares – souvent extrêmement rares – qui ont tendance à se réunir dans des lieux particuliers. Par exemple, Kanouse, Berry et Duan (1999) ont employé la méthode pour échantillonner les prostituées de la rue dans le comté de Los Angeles en échantillonnant les endroits où l'on sait qu'a lieu la prostitution de rue, ainsi que des périodes (jours et quarts durant les jours). L'échantillonnage de lieux de sociabilité (ou de centres) a également été utilisé pour échantillonner les immigrants légaux et illégaux en Italie (Mecca 2004). Dans le cadre d'une enquête auprès de la population d'immigrants établis à Milan réalisée en 2002, 13 types de centres ont été relevés, variant des centres qui fournissent des listes partielles d'après des sources admistratives (par exemple centres juridiques et de travail, cours de langue) à des centres fournissant des renseignements des personnes qui les fréquentent (centres de services de bien-être, associations culturelles) et à des centres ne possédant aucune information sur la base de sondage (par exemple centres commerciaux, boutiques ethniques).

organismes des CPS et tout autre organisme. Le plan d'échantillonnage a par conséquent été traité comme un plan à base de sondage double, où les CPS représentaient l'une des bases de sondage et les autres bases de sondage combinées, la deuxième base de sondage (c'est-à-dire en supposant qu'il n'existait aucun chevauchement entre les autres bases de sondage).

3.5 Échantillonnage par réseaux

L'échantillonnage par réseaux (ou basé sur la multiplicité) étend l'approche classique de présélection en demandant aux personnes (ou adresses) échantillonnées de servir également de répondants par procuration pour fournir l'information de présélection pour les personnes avec lesquelles ils sont liés d'une manière spécifiée clairement (Sudman et coll. 1988 ; Sirkén 2004, 2005). Des personnes appartenées, telles que les parents, les frères et sœurs et les enfants servent souvent de base pour l'établissement des liens. Une exigence clé est que chaque membre du lien doit connaître et être disposé à communiquer le statut d'appartenance à la population rare de toutes les personnes avec lesquelles il a un lien. Dans une étude pilote des anciens combattants de la guerre du Vietnam de sexe masculin, Rothbart, Fine et Sudman (1982) ont inclus les oncles et les tantes comme répondants au même titre que les parents et les frères et sœurs, mais ont constaté que les oncles et les tantes identifiaient un nombre nettement plus faible que prévu d'anciens combattants de la guerre du Vietnam. Le fait que les oncles et les tantes semblent ne pas déclarer certains anciens combattants cause un biais d'échantillonnage éventuel, si bien que les liens dans les règles de jumelage est problématique.

Les pistes multiples de sélection qui existent dans l'échantillonnage par réseaux doivent être prises en compte dans la détermination des probabilités de sélection de la même manière que celle décrite pour les bases de sondage multiples à la section précédente. Conceptuellement, on peut considérer que chaque membre de la population rare est divisé en, disons, l parties correspondant aux l répondants pour le membre en question ; ce sont ensuite ces parties qui sont échantillonnées pour l'enquête. Voir Lavalée (2007) pour un exposé sur la théorie qui sous-tend la méthode. Quand l'échantillonnage par réseaux est utilisé dans les enquêtes conçues pour recueillir des données sur les caractéristiques des membres d'une population rare, une prise de contact directe doit avoir lieu avec les membres de la population rare identifiés par le répondant initial. Dans ce cas, ce dernier a dû pouvoir fournir les coordonnées des membres de la population rare. La définition des liens peut être structurée de manière à faciliter la collecte des données, par exemple, dans le cas de l'interview sur place, les liens peuvent être limités aux personnes appartenées

Certaines formes d'établissement de liens ont l'avantage supplémentaire de permettre d'intégrer des membres de la population rare qui ne figurent pas dans la base d'échantillonnage originale et qui, par conséquent, représenteraient autrement une composante de la non-couverture. Par exemple, Brick (1990) décrit un essai sur le terrain pour la National Household Education Survey (NHES) réalisé par téléphone en recourant à l'échantillonnage par réseaux pour sélectionner l'échantillon du groupe des 14 à 21 ans, en se concentrant sur les jeunes ayant abandonné leurs études. Dans un sous-échantillon de ménages, on a demandé à toutes les femmes de 28 à 65 ans de fournir des renseignements sur tous les enfants de 14 à 21 ans vivant ailleurs au moment de l'entrevue. Certains de ces enfants vivaient dans des ménages possédant le téléphone si bien que leur sélection pouvait se faire selon deux pistes. D'autres vivaient dans des ménages n'ayant pas le téléphone et n'auraient donc pas été couverts par l'enquête ; leur inclusion au moyen de l'échantillonnage par réseaux a augmenté le taux de couverture d'environ 5 % en 1989. Toutefois, le taux de réponse pour les jeunes vivant hors du ménage était nettement plus faible que pour ceux vivant dans le ménage,

vivant dans une zone définie à proximité de la personne qui fournit l'information. Sudman et Freeman (1988) décrivent l'application de l'échantillonnage par réseaux dans une enquête téléphonique au sujet de l'accès aux soins de santé pour laquelle il a fallu suréchantillonner les personnes ayant une maladie chronique ou grave. Au cours d'une première prise de contact avec le chef du ménage, les liens avec les parents, beaux-parents, frères et sœurs, grands-parents et petits-enfants de moins de 18 ans du répondant ou de son (sa) conjoint(e) ont été déterminés et les données ont été recueillies sur l'état de santé de ces personnes. L'utilisation de ce plan d'échantillonnage par réseaux a permis d'accroître le nombre identifié d'adultes atteints d'une maladie chronique ou grave d'environ un tiers. Cependant, environ un répondant sur huit du réseau initial ayant des personnes appartenées ne pouvait ou ne voulait pas fournir de renseignements sur les maladies des membres de leur réseau, et 70 % n'ont pas fourni de coordonnées complètes, y compris 28 % qui n'ont fourni ni le nom ni les coordonnées (ce qui rendait le dépistage impossible). L'utilisation de l'échantillonnage par réseaux a donné lieu à certains résultats faussés positifs (personnes déclarées comme étant atteintes d'une maladie chronique ou grave par le répondant initial, mais déclarant elle-même bien se porter). Une pré-occupation plus grave est que l'enquête n'a pas pu fournir l'information sur les résultats faussés négatifs (il aurait fallu pour cela suivre un échantillon des membres des réseaux qui, selon les déclarations du répondant initial, se portaient bien).

classes de pondération pour la non-réponse qui tiennent compte de l'appartenance à d'autres bases de sondage. Les concepteurs d'enquêtes doivent plutôt supposer que, dans les classes de pondération, les taux de réponse sont les mêmes quel que soit le nombre de bases de sondage dans laquelle figure l'unité.

En général, pour appliquer l'approche qui vient d'être décrite, il faut connaître les probabilités de sélection de chaque unité échantillonnée pour toutes les bases de sondage, information qui n'est pas toujours disponible. Quand on ne connaît pas les probabilités de sélection pour d'autres bases de sondage (que celles) dont est tirée l'unité (mais que la présence/absence dans les bases de sondage est connue), on peut recourir à une autre approche, appelée méthode du partage des poids élaborée par Lavallée (1995, 2007). Des estimations sans biais des totaux de population sont obtenues si le poids appliqué à l'unité i est donné par $w_i = \sum_j \alpha_j w_{ij}$, où α_j représente tout ensemble de constantes tel que $\sum_j \alpha_j = 1$ quand la sommation est effectuée sur les j bases de sondage, $w_{ij} = 1/p_j$ si l'unité i est sélectionnée dans la base de sondage j avec la probabilité p_j et $w_{ij} = 0$ autrement (Kallion et Brick 1995, Lavallée 2007). Pour de nombreuses applications, il est raisonnable de poser que $\alpha_j = \alpha_j$ et un bon choix de α_j est alors $\alpha_j = n_j / \sum_j n_j$, où n_j est la taille d'échantillon efficace fondée sur un effet de plan moyen (Chu, Brick et Kallion 1999).

La deuxième approche générale pour traiter les pistes multiples de sélection s'appuie sur la méthodologie des bases multiples introduites par Hartley (1974) et a été le sujet de nombreux travaux de recherche récents (voir, par exemple, Lohr et Rao 2000 et 2006, ainsi que les références mentionnées dans ces articles). Dans le cas de deux bases de sondage (A et B), la population peut être divisée en trois sous-ensembles mutuellement exclusifs étiquetés $a = A \cap B$, $b = \bar{A} \cap B$ et $ab = A \cap \bar{B}$. L'échantillon peut être divisé en échantillons provenant de a , de b et de ab , où l'échantillon ab peut être réparti en un groupe de répondants échantillonnés dans la base de sondage A et un groupe échantillonné dans la base B . Pour les échantillons dans les sous-ensembles a et b , il n'existe qu'une seule piste de sélection de sorte qu'ils sont traités facilement dans l'estimation. Les totaux pour ab pourraient être estimés en se basant sur l'échantillon provenant de la base de sondage A ou celui provenant de la base de sondage B , disons \bar{Y}_{ab}^A . La méthode de Hartley consiste à calculer la moyenne pondérée de ces deux estimateurs, $\bar{Y}_{ab} = \theta \bar{Y}_{ab}^A + (1 - \theta) \bar{Y}_{ab}^B$, où θ est choisi pour minimiser la variance de \bar{Y}_{ab} , comprise dans l'échantillon. Notons que la méthodologie des bases de sondage doubles dépend de l'estimateur, les valeurs de θ étant différentes pour des estimateurs différents, Skinner et Rao (1996), ainsi que

poids de Lavallée décrite plus haut.

Quand on utilise un plan à base de sondage double ou multiple, il arrive souvent que la couverture d'une des bases de sondage soit complète, mais que la prévalence de la population rare y soit faible (par exemple, une base aréolaire) et que la prévalence de la population rare dans une ou plusieurs autres bases de sondage soit beaucoup plus élevée, mais que sa couverture soit incomplète. Ainsi, Metcalfe et Scott (2009) ont combiné un échantillon aréolaire avec un échantillon de listes électorales pour l'Australia and Diabetes Heart and Health Survey, dans laquelle les personnes âgées étaient des îles du Pacifique, les Maoris et les personnes âgées étaient des domaines d'intérêt particulier. La base des listes électorales avait l'avantage de contenir des renseignements sur l'âge des électeurs, ainsi qu'une liste spéciale sur laquelle les personnes qui se considéraient personnellement d'ascendance maorie pouvaient s'inscrire. En outre, de nombreuses personnes originaires du Pacifique pouvaient vraisemblablement être identifiées d'après leur nom, puisque les langues du Pacifique comprennent moins de lettres que l'anglais. Un échantillon stratifié disproportionnellement a été sélectionné dans la base des listes électorales pour suréchantillonner les domaines d'intérêt et l'échantillon tiré de la base aréolaire a fourni les personnes qui ne figuraient pas sur les listes électorales.

La National Incidence Study of Child Abuse and Neglect offre un exemple d'une situation plus complexe (Wingate, Park, Rust, Liu et Shapiro 2007). Cette enquête s'est appuyée sur plusieurs bases de sondage pour accroître la couverture globale des enfants maltraités et négligés. Les bureaux des Child Protective Services (CPS) compris dans la base de sondage principale, tandis que les bureaux de police, les hôpitaux, les écoles, les réfugiés, les garderies et d'autres organismes étaient les sources d'autres bases de sondage. Les échantillons de bureaux des CPS ont été sélectionnés d'après des listes, mais les échantillons des autres organismes ont été tirés en échantillonnant les organismes, en dressant des listes du personnel professionnel pertinent et en échantillonnant des membres du personnel qui ont joué le rôle de répondants au sujet des enfants maltraités. En suivant ces procédures, le chevauchement entre les organismes ne peut pas être déterminé, sauf celui entre les

3.4 Bases de sondage multiples

Kalton : Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales

les strates, ce qui permet de traiter les listes comme étant en blanc (Kish 1965b, pages 132 à 139), et de faire l'analyse par les méthodes classiques.

L'utilisation de la méthode d'identification unique peut parfois être inefficace s'il faut communiquer avec les personnes échantillonnées à partir d'une base de sondage en vue de déterminer si leur listage doit être traité comme réel ou comme un blanc pour la base de sondage en question. Dans ces conditions, il est généralement plus économique de recueillir les données d'enquête pour les personnes échantillonnées (c'est-à-dire accepter les pistes multiples de sélection). Il existe cependant des exceptions, comme dans le cas de la National Survey of America's Families. Cette enquête s'appuie sur la combinaison d'une base aréolaire et d'une liste de numéros de téléphone obtenus par C.A., la base aréolaire étant utilisée pour couvrir uniquement les ménages sans téléphone (Waksberg, Brick et coll. 1997). Cette approche s'est avérée efficace pour effectuer une pré-sélection rapide des ménages présents dans la base aréolaire afin d'éliminer ceux possédant le téléphone et de retenir uniquement ceux sans téléphone pour l'enquête.

Deux approches générales permettent de tenir compte des pistes multiples de sélection dans le calcul des probabilités de sélection (Bankier 1986; Kalton et Anderson 1986). Une méthode consiste à calculer la probabilité globale de sélection de l'unité échantillonnée sur toutes les bases de sondage et d'utiliser l'inverse de cette probabilité comme poids de base pour l'analyse (ce qui mène à l'estimateur d'Horvitz-Thompson). Par exemple, la probabilité globale de sélection de l'unité échantillonnée i dans deux bases de sondage est $p_i = (p_{1i} + p_{2i} - p_{1i}p_{2i})$, où p_{1i} est la probabilité de sélection de l'unité dans la base $f = 1, 2$. Une variante consiste à remplacer la probabilité globale de sélection par le nombre prévu de sélections (ce qui mène à l'estimateur de Hansen-Hurwitz), qui est plus facile à calculer quand on se sert de plusieurs bases de sondage. S'il n'existe que deux bases de sondage, le nombre prévu de sélections est $(p_{1i} + p_{2i})$. Quand les probabilités de sélection sont faibles, ces deux estimateurs diffèrent peu.

Les corrections pour tenir compte de la non-réponse et pour caler les totaux d'échantillon sur les totaux connus de population peuvent être appliquées aux probabilités globales de sélection p_i ou individuellement aux p_{fj} . L'un des problèmes est que les concepteurs d'enquête ne savent pas si une unité non répondante échantillonnée dans une base de sondage figure dans une autre base de sondage, puisque cette information n'est recueillie que durant l'interview. Le cas échéant, les p_i pour les unités non répondantes ne peuvent pas être calculées directement et doivent être estimées d'une certaine façon. Quand les corrections sont appliquées aux unités échantillonnées représentées des sous-classes dans

Parfois, on dispose de bases de sondage plus axées sur une population rare qu'une base de sondage générale, mais qui ne couvrent qu'une partie de cette population rare. Le cas échéant, il peut être efficace de sélectionner l'échantillon à partir de plus d'une base de sondage. Par exemple, dans le cas fréquent du suréchantillonnage des minorités ethniques, on dispose parfois d'une liste. Les personnes figurant sur la liste peuvent être classées en fonction de leur nom comme appartenant probablement à un groupe ethnique donné (par exemple, Chinois, Coréens, originaires des îles du Pacifique, Vietnamiens) pour créer une deuxième base de sondage incomplète dont on peut tirer un échantillon en plus de celui sélectionné dans la base de sondage plus complète dans laquelle la prévalence de la population rare est plus faible (voir, par exemple, Elliott et coll. 2008; Flores Cervantes et Kalton 2008). Comme dans le cas de la stratification disproportionnée (section 3.2), des avantages importants ne peuvent être tirés de cette approche que si la deuxième base de sondage offre une prévalence élevée et couvre une fraction appréciable de la population rare. Consulter Lohr (2009) pour une revue des questions relatives à l'échantillonnage à partir de bases de sondage multiples.

Si les bases de sondage sont multiples, certains membres de la population rare pourraient être compris dans plusieurs d'entre elles, auquel cas il pourrait exister plusieurs pistes pour les sélectionner dans l'échantillon. Trois grandes approches existent pour traiter ces cas de multiplicité (Anderson et Kalton 1990; Kalton et Anderson 1986). Quand toutes les bases de sondage sont des listes, ce qui est parfois le cas dans les études de la santé, il peut être possible de les combiner en une seule liste sans enregistrements en double; cependant, cette fusion des listes pose souvent des problèmes compliqués de couplage d'enregistrements. Note approche consiste à rendre les listes non chevauchantes en utilisant une règle d'identification unique qui associe chaque membre de la population rare à une seule base de sondage et en traitant les listes des autres bases de sondage comme des listes en blanc (Kish 1965b, pages 388 à 390). Les échantillons sont tirés de chacune des bases de sondage sans tenir compte des enregistrements en double, mais seules les listes échantillonnées non en blanc sont acceptées pour l'échantillon final. Cette approche donne de meilleurs résultats quand des recherches peuvent être effectuées pour chaque unité échantillonnée dans les autres bases de sondage; si les bases de sondage sont classées par ordre de priorité et que l'unité est découverte dans une base de sondage qui précède celle à partir de laquelle la sélection a été faite, la liste échantillonnée est considérée comme une liste en blanc. Dans ces conditions, les bases de sondage sont des strates et les unités échantillonnées représentent des sous-classes dans

conçue pour recueillir une vaste gamme de mesures sur les répondants échantillonnés, y compris une batterie de mesures cognitives. En se basant sur ces mesures, on a réparti les répondants à la HRS en cinq strates cognitives, puis on a sélectionné un échantillon disproportionnellement stratifié à la deuxième phase. La collecte coûteuse des données de deuxième phase comprenait une évaluation à domicile structurée d'une durée de trois à quatre heures effectuée par une infirmière et un technicien en neuropsychologie. Les résultats de l'évaluation ont été évalués par un gériopsychiatre, un neurologue et un spécialiste de la neurosciences cognitive en vue d'attribuer un diagnostic provisoire d'état cognitif qui a été réévalué par la suite à la lumière des données figurant dans les dossiers médicaux des sujets (Langa, Plassman, Wallace, et coll. 2005).

Un troisième exemple est celui d'un plan d'échantillonnage à trois phases qui a été utilisé dans une étude pilote en vue de repérer les personnes qui auraient droit aux prestations d'incapacité de la Social Security Administration des États-Unis si elles en faisaient la demande (Maffeo, Frey et Kalton 2000). En premier lieu, on a demandé à un membre bien informé du ménage de fournir des renseignements au sujet de la situation de bénéficiaire de prestations d'incapacité et de la situation d'invalidité de tous les adultes de 18 à 69 ans faisant partie du ménage. À la deuxième phase, toutes les personnes classées dans une strate de non-bénéficiaires gravement handicapés et des échantillons des autres strates ont été interviewés sur place, puis reclassifiés au besoin dans des strates d'incapacité probable pour la troisième phase. Durant cette dernière, on a sélectionné un échantillon disproportionnellement stratifié de personnes en vue de leur faire subir des examens médicaux dans des centres d'examen mobiles.

Dans les plans d'échantillonnage à deux phases, une pratique relativement courante consiste à ne pas tirer d'échantillon de deuxième (ou de troisième) phase dans la strate de sujets classés comme étant des non-membres du domaine rare en fonction des réponses qu'ils ont données à la phase précédente. La proportion de la population comprise dans cette strate est habituellement très élevée et la prévalence du domaine rare y est faible (en fait, comme dans l'étude de Haerter, Anderson et Schoenberg (1986), la strate est souvent définie prudemment dans le but d'éviter l'inclusion des sujets qui pourraient être membres du domaine rare). Par conséquent, un échantillon de cette strate de taille modérée ne produira presque aucun membre du domaine rare. Toutefois, la stratégie consistant à ne tirer aucun échantillon de cette strate est risquée. Si la prévalence du domaine rare dans cette grande strate est plus que minime, une proportion importante du domaine risque de ne pas être représentée dans l'échantillon final.

un mode d'inférence fondé sur un modèle dans lequel les poids d'échantillonnage sont ignorés. À mon avis, ne pas tenir compte des poids d'échantillonnage pose problème. Cependant, la discussion de cette question dépasse le cadre du présent article.

3.3 Échantillonnage à deux phases

L'approche de présélection traitée aux sections 3.1 et 3.2 repose sur l'hypothèse que l'identification des membres d'une population rare est relativement aisée. Si l'identification exacte est coûteuse, il peut être utile d'adopter un plan d'échantillonnage à deux phases, en commençant par une classification de présélection imparfaite à la première phase, suivie par l'identification exacte d'un sous-échantillon stratifié disproportionnellement à la deuxième phase. La rentabilité de l'approche à deux phases dépend en partie des coûts relatifs de la classification imparfaite et de l'identification exacte : puisque la classification imparfaite consomme une partie des ressources de l'étude, elle doit être nettement moins coûteuse que l'identification exacte. Deming (1977) a suggéré que le ratio des coûts par unité entre la collecte de données de deuxième phase et celle de première phase soit au moins de 6 pour 1. En outre, la classification imparfaite doit être raisonnablement efficace afin de tirer des avantages importants d'une stratification de deuxième phase disproportionnée.

L'échantillonnage à deux phases, voire même à trois, est souvent utile dans les enquêtes médicales auprès de personnes ayant des problèmes de santé particuliers. Fréquemment, la première phase de l'enquête correspond à l'administration d'un questionnaire de présélection par les intervieweurs, tandis que la deuxième phase est effectuée par des cliniciens, souvent dans un centre médical. Ainsi, dans une enquête sur l'épilepsie dans le comté de Copiah, au Mississippi, Haerter, Anderson et Schoenberg (1986) ont d'abord recouru à des intervieweurs pour administrer à tous les ménages du comté un questionnaire qui avait été prétesté pour s'assurer qu'il permettait de déprimer les personnes atteintes d'épilepsie avec un haut niveau de sensibilité. Afin d'éviter les résultats faussement négatifs durant cette première phase, un grand réseau de présélection a été utilisé pour repérer les personnes qui participeraient à la deuxième phase. Toutes celles qui ont été identifiées ont été les sujets de la deuxième phase de l'enquête qui consistait en de brefs examens neurologiques effectués par une équipe de quatre neurologues principaux dans une clinique de santé publique. Un deuxième exemple illustre l'utilisation d'une autre enquête comme collecte des données de première phase pour l'étude d'un domaine rare. Dans l'exemple, la Health and Retirement Study (HRS) a été utilisée à la première phase d'une étude de la démence et d'autres troubles cognitifs chez les adultes de 70 ans et plus. La HRS est

adultes bouddhistes. Les résidences affectées à la strate « probable » (environ 20 %) ont ensuite été échantillonnées à un taux quatre fois plus élevé que les autres.

Parfois, si l'enquête est destinée à ne produire des estimations que pour une population très rare, la stratification disproportionnée peut encore nécessiter une présélection excessive. Le cas échéant, l'échantillonnage doit parfois se faire dans les strates où la prévalence est la plus élevée, en laissant tomber les autres et en acceptant un certain degré de non-couverture (ou en redéfinissant la population étudiée afin d'inclure uniquement les membres de la population rare compris dans les strates qui ont été échantillonnées). La (HHANES) de 1982-1994 en est un exemple. Pour obtenir les échantillons de Mexicano-américains dans le Sud-Ouest et de Portoricains dans la région de la ville de New York, l'échantillonnage a été effectué uniquement dans les comtés où les nombres et (ou) pourcentages d'Hispaniques, basés sur les chiffres du Recensement de 1980, étaient élevés (Gonzalez, Ezzati, White, Massey, Lago et Waksberg 1985).

À titre d'autre exemple de cette approche, Hedges (1979) décrit une méthode pour échantillonner une population minoritaire qui est plus concentrée dans certains districts géographiques, tels que les districts de recensement, que dans d'autres. Dans cette méthode, les districts sont énumérés par ordre de prévalence des membres de la population rare (obtenue, disons, d'après le dernier recensement), puis les concepteurs d'enquête produisent des courbes de Lorenz de la distribution cumulative de la prévalence de la population rare et de la distribution cumulative des portions de membres de la population rare couverte. La prévalence cumulative diminuant à mesure que la couverture cumulée augmente, les concepteurs peuvent utiliser cette distribution pour choisir la combinaison de strates et de proportions couverte qui répond le mieux à leurs exigences. La question est alors de savoir s'il faut faire les ajustements pour la population couverte ou pour la population complète en corrigeant la pondération de la population pour essayer de tenir compte du biais de non-

couverture. Quand un domaine est très rare mais qu'une partie de celui-ci est fortement concentrée dans une strate, les chercheurs échantillonnent parfois cette strate à un taux nettement plus élevé que la valeur optimale afin de produire un nombre appréciable de cas. Bien que cette approche puisse produire un grand échantillon de la population rare, la taille efficace d'échantillon (c'est-à-dire la taille d'échantillon divisée par l'effet de plan) sera plus faible que si les fractions d'échantillonnage optimales avaient été utilisées. Donc, du point de vue du mode d'inférence classique fondé sur le plan de sondage, cette approche ne convient pas. Toutefois, les chercheurs qui l'utilisent défendent souvent

peuvent être un moyen raisonnablement efficace de repérer les Hispaniques, les Philippins, les Vietnamiens, les Japonais et les Chinois, mais non les Noirs. Un certain nombre de listes de noms associés à différents groupes raciaux/ethniques peuvent être dressées, telle la liste des noms espagnols produite par le U.S. Census Bureau pour les années 1990 (Word et Perkins 1996). Plusieurs fournisseurs de logiciels commerciaux ont développé des algorithmes complexes pour exécuter des classifications raciales/ethniques fondées sur les noms (voir Fiscella et Fremont 2006 pour plus de renseignements). L'utilisation des noms pour déterminer la race et l'ethnicité a suscité beaucoup d'intérêt chez les épidémiologistes et les démographes, qui ont procédé à un certain nombre d'évaluations de la méthode (par exemple Landerdale et Kestenbaum 2000; Elliott, Morrison, Fremont, McCaffrey, Pantofa et Lurie 2009). Ces chercheurs évaluent souvent l'efficacité de la méthode pour ce qui est de sa valeur prédictive positive et de sa sensibilité, qui sont les équivalents de la prévalence et de la proportion de membres du domaine qui sont identifiés en tant que tels par l'instrument utilisé pour la classification. Dans le contexte de l'échantillonnage, outre les limites de l'instrument, les chercheurs doivent aussi tenir compte du fait que, parfois, les noms ne sont pas disponibles et que certains de ceux disponibles peuvent être incorrects (par exemple, dans le cas de l'échantillonnage fondé sur l'adresse, les noms pourraient ne plus être à jour, parce que la famille originale a déménagé et qu'une nouvelle famille s'est installée à l'adresse en question). Ces contraintes supplémentaires réduisent l'efficacité de la stratification en fonction du nom et, selon les circonstances, la réduction peut être appréciable.

Comme pour la stratification en général, les facteurs de stratification choisis pour échantillonner les populations rares ne doivent pas se limiter à des mesures objectives. Il peut fort bien s'agir de classifications subjectives. Le seul point à prendre en considération est la mesure dans laquelle les facteurs répondent aux besoins de la stratification (voir Kish 1965b, pages 412 à 415, pour un exemple de l'efficacité de l'utilisation de la classification rapide des logements du programme de traitement de listes dans les catégories de statut socioéconomique faible, moyen ou élevé pour la stratification disproportionnée). Elliott, McCaffrey, Perlman, Marshall et Hambarsoomians (2009) décrivent une application efficace de la stratification sub-jetive pour l'échantillonnage des immigrants cambodgiens à Long Beach, en Californie. Un expert de la communauté locale a déterminé, pour toutes les résidences individuelles dans les îlots échantillonnés, s'il était probable ou improbable qu'elles contiennent un ménage cambodgien en se basant sur des caractéristiques culturelles observables extérieurement, telles que les chaussures devant la porte et les

de la collecte principale des données par unité échantillonnée ne doit pas être élevée. Souvent, ces conditions ne sont pas toutes trois satisfaites, auquel cas les gains sont modestes. De surcroît, les résultats susmentionnés sont fondés sur l'hypothèse que la prévalence réelle de la population rare dans chaque strate est connue, alors qu'en pratique, les chiffres disponibles seront dépassés (par exemple fondés sur le dernier recensement) ou peut-être tout simplement calculés approximativement. Les erreurs dans les estimations de la prévalence réduisent les gains de précision réalisés grâce à la stratification disproportionnée et pourraient même entraîner une perte de précision. Une surestimation importante de la prévalence de la population rare et, donc, de la fraction d'échantillonnage optimale dans la strate à forte densité peut réduire considérablement la précision des estimations fondées sur l'enquête. Il est donc souvent préférable d'adopter une stratégie prudente, autrement dit de procéder à une répartition un peu moins disproportionnée, plus proche de la répartition proportionnelle.

3.2.2 Applications

En cas d'échantillonnage aréolaire, les données provenant du dernier recensement et d'autres sources peuvent être utilisées pour répartir les grappes aréolaires entre les strates en fonction de leur prévalence estimée pour la population rare. Voir Waksberg, Judkins et Massey (1997) pour une étude détaillée de cette approche en vue de suréchantillonner diverses populations raciales/ethniques et la population à faible revenu en prenant pour grappes les îlots et les groupes d'îlots du recensement des États-Unis. En se fondant sur les données du Recensement de 1990, Waksberg et ses collègues ont constaté que l'approche donnait généralement de bons résultats pour les Noirs et les Hispaniques, mais non pour la population à faible revenu. Même si la concentration de la population à faible revenu était élevée dans certains blocs et groupes d'îlots, ces derniers ne couvraient pas une forte proportion de cette population.

Quand les concepteurs d'enquêtes ont accès à une liste des strates de membres probables de certains groupes raciaux/ethniques. Cette situation se produit, par exemple, lorsque l'on dispose de listes de noms et de numéros de téléphone et que les noms sont fusionnés avec les adresses du Delivery Sequence File du service postal des États-Unis (USPS) (aucune fusion de noms n'est faite dans certains cas). La répartition entre les strates peut être fondée sur les noms de famille uniquement ou sur une combinaison du nom de famille et du premier prénom (et même d'autres noms). Puisque les femmes adoptent souvent le nom de famille de leur mari, la répartition est généralement plus meilleure pour les hommes que pour les femmes. Les noms

échantillonnés de la population rare au coût pour un membre suit donne le ratio de la variance de la moyenne d'échantillon quand les fractions d'échantillonnage stratifié disproportionné sont optimales à celle sous échantillonnage stratifié proportionnel de même coût total :

$$R = \frac{\left[\sum A_h \sqrt{P(c-1) + \frac{P}{P_h}} \right]^2}{P(c-1) + 1}$$

où A_h est la proportion de la population rare dans la strate h et P est la prévalence de la population rare dans la population complète. En général, la variabilité inter-strates des fractions d'échantillonnage optimales et les gains de précision pour la moyenne d'échantillon diminuent à mesure que c augmente. Donc, si le coût de la collecte principale des données d'enquête est élevé, comme cela est le cas quand l'enquête comprend un examen médical approfondi, ou que le coût de la présélection est très faible, la stratification disproportionnée ne produit parfois que de faibles gains de précision.

Si le coût de la collecte principale des données n'ajoute rien au coût de présélection, le ratio du coût de la collecte des données principale au coût de la présélection sera égal à $c = 1$. Dans cette situation limite, les formules susmentionnées se simplifient en $f_h \propto \sqrt{P_h}$ et $R = (2\sqrt{A_h W_h})^2$, où W_h est la proportion de la population totale dans la strate h . Ces formules simples donnent une bonne idée de la variation maximale des fractions d'échantillonnage optimales et des gains maximaux de précision qui peuvent être réalisés. Dans la formule de la fraction d'échantillonnage optimale, la fonction racine carrée indique clairement que les prévalences dans les strates doivent varier considérablement pour que les fractions d'échantillonnage s'écartent de manière appréciable de la répartition proportionnelle. Par exemple, même si la prévalence dans la strate A est quatre fois plus élevée que dans la strate B , la fraction d'échantillonnage optimale dans la strate A n'est que deux fois plus grande que celle dans la strate B . Les gains de précision $(1-R)$ sont importants si A_h est grande quand W_h est faible et inversement. S'il n'existe que deux strates, une strate dans laquelle la prévalence est cinq fois plus élevée que la prévalence globale (c -est-à-dire $P_h/P = 5$) produira un gain de précision de 25 % ou plus ($(1-R) \geq 0,25$) uniquement si elle comprend au moins 60 % de la population rare (Kallion 2003, tableau 1). En bref, quoique généralement utile, la stratification disproportionnée ne produit des gains d'efficacité appréciables que si trois conditions sont remplies : 1) la prévalence de la population rare doit être beaucoup plus élevée dans la strate suréchantillonnée, 2) la strate suréchantillonnée doit contenir une forte proportion de la population rare et 3) le coût

ces questions produira une erreur de classification (Sudman 1972, 1976). Quand on procède à un suréchantillonnage d'un ou de plusieurs domaines rares dans le cadre d'une enquête auprès de la population générale, les erreurs de classification sont de deuxième phase, ce qui évite la non-couverture. Les erreurs de classification produisent néanmoins des tailles d'échantillon plus petites pour les domaines rares ; de surcroît, la variation des poids d'échantillonnage entre les répondants sélectionnés comme membres du domaine rare et ceux échantillonnés comme membres d'un autre domaine peut entraîner une perte considérable de précision. La non-couverture est plus problématique quand les données de présélection sont recueillies auprès de répondants par procuration. Le problème est particulièrement épineux dans le cas du dénombrement ciblé.

Dans un certain nombre d'enquêtes sur des populations rares, la proportion de membres identifiés de la population rare était nettement plus faible que les taux de prévalence de référence. Par exemple, dans le cas de la NIS de 1994, le défaut dans la proportion identifiée d'enfants de 19 à 35 mois était considérable (4,1 % comparativement au taux prévu de 5 %) (Camburn et Wright 1996). Dans la National Longitudinal Survey of Youth de 1997, 75 % seulement des jeunes de 12 à 23 ans ont été repérés (Horrigan, Moore, Pedlow et Wolter 1999). Ces constatations pourraient résulter de taux de réponse plus élevés pour les membres de la population rare, de divers types de non-couverture de la base de sondage ou d'erreurs de classification en ce qui concerne l'appartenance au domaine. Pour produire la taille d'échantillon requise, une allocation doit être faite pour la sous-représentation à l'étape de l'allocation du plan d'échantillonnage. La non-couverture d'un domaine d'âge peut-être parce que les répondants ne connaissent pas les âges exacts (ceux incorrectement rejetés à la présélection étant perdus et ceux incorrectement retenus à la présélection étant détectés et éliminés plus tard) ou parce que les répondants font délibérément une déclaration incorrecte pour éviter l'interview de suivi. Pour contourner cet effet, il peut être utile de commencer par une présélection initiale portant sur tous les membres du ménage ou sur un groupe d'âge plus étendu, puis de réduire la présélection au groupe d'âge requis par après.

Des corrections de la pondération peuvent être faites pour essayer d'atténuer les biais causés par la non-réponse et la non-couverture, mais elles sont forcément imparfaites. La correction pour tenir compte du niveau de non-réponse propre à un domaine requiert que l'on sache si les non-répondants appartiennent ou non au domaine, mais souvent, ces renseignements n'existent pas. Les corrections pour la non-couverture d'un domaine rare nécessitent des données externes précises pour le domaine, données qui souvent ne

3.2 Stratification disproportionnée

Une extension naturelle de l'approche de présélection consiste à déterminer dans quelles strates la présélection sera la plus productive. Le cas idéal est celui où une ou plusieurs strates englobant la totalité de la population rare mais aucune unité extérieure à cette population sont identifiées. Aucun processus de présélection n'est alors nécessaire. En dehors de ces conditions idéales, des échantillons doivent être sélectionnés dans toutes les strates (à part celles que l'on sait ne contenir aucun membre de la population rare) pour arriver à une couverture complète de la population rare. L'utilisation de la stratification disproportionnée, dans laquelle les fractions d'échantillonnage sont plus élevées dans les strates où la prévalence de la population rare est plus élevée, permet de réduire la

3.2.1 Contexte théorique

Considérons au départ une enquête conçue pour fournir des estimations pour une seule population rare. Waksberg (1973) a effectué une première évaluation théorique de l'utilité de la stratification disproportionnée dans ces conditions. Les articles subséquents traitant de ce sujet comprennent ceux de Kalton et Anderson (1986) et Kalton (1993a, 2003). L'objectif de la stratification disproportionnée pour l'échantillonnage d'une seule population rare, à savoir le taux de prévalence dans chaque strate, la proportion de la population rare dans chaque strate et le ratio du coût total de la collecte des données pour les membres des populations rare et non rare sont les mêmes dans toutes les strates et que 2) les coûts de la collecte des données élémentaires pour la population rare sont les mêmes dans cette population. Si l'on suppose que 1) les variances de la présélection effectuée pour identifier les membres des données pour les membres de la population rare au coût des données pour les membres de la population rare sont les mêmes dans toutes les strates et que 2) les coûts de la collecte des données pour les membres des populations rare et non rare sont les mêmes dans toutes les strates, alors, sous échantillonnage aléatoire simple dans les strates, la fraction d'échantillonnage optimale dans la strate h pour minimiser la variance d'une moyenne estimée pour la population rare, sous la contrainte d'un budget total fixe, est donné par

$$f_h \propto \sqrt{\frac{P_h(c-1)+1}{P_h}}$$

où P_h est la proportion des unités présentes dans la strate h qui sont membres de la population rare et c est le ratio du coût de la collecte des données pour un membre

questionnaire de présélection n'est pas obtenue ou si un membre de la population rare est identifié (peut-être par un répondant par procuration), mais ne répond pas aux questions de l'enquête. Le taux global de non-réponse pourrait fort bien être nettement plus élevé qu'il ne le serait sans la composante de présélection. En outre, les concepteurs d'enquêtes doivent tenir compte de la nature du domaine rare et de la façon dont les membres de ce domaine réagissent au contenu de l'enquête. Une enquête dans laquelle on demande à de nouveaux immigrants des renseignements sur leur expérience d'immigration pourrait produire un taux de réponse fort différent d'une enquête auprès d'anciens combattants au sujet des services médicaux et autres services de soutien qu'ils reçoivent.

Le troisième problème est que la non-couverture peut être importante en cas de présélection à grande échelle pour repérer les populations rares. L'une des sources de non-couverture à trait à la base de sondage utilisée pour l'échantillon de présélection. Même si la couverture globale d'une base de sondage est bonne, sa couverture d'un domaine rare peut être inadéquate. Ainsi, la non-couverture d'une liste de numéros de téléphone conventionnel est beaucoup plus importante pour les jeunes ménages que pour l'ensemble de la population. Les concepteurs d'enquêtes par téléphone conventionnel portant sur des domaines rares tels que les jeunes enfants et les étudiants collégiens doivent donc examiner attentivement la possibilité de biais de non-couverture. Afin de résoudre le problème de la non-couverture importante des personnes pauvres dans les enquêtes téléphoniques, la National Survey of America's Families, qui a été conçue pour observer le bien-être des enfants et des adultes à la suite de réformes du régime de bien-être, comportait un échantillon aréolaire de ménages sans téléphone en plus de l'échantillon principal de numéros de téléphone sélectionnés par la méthode de composition aléatoire (CA) (Waksberg, Brick, Shapiro, Flores Cervantes et Bell 1997).

Une autre source de non-couverture résulte du fait que certains membres de la population rare ne sont pas identifiés à l'étape de la présélection. En particulier, si une enquête a pour objectif de recueillir des données uniquement après des membres d'un domaine rare, certains répondants de la phase de présélection pourraient déclarer faussement, et certains intervieweurs pourraient enregistrer faussement, que le domaine. Ces erreurs de classification peuvent être commises par inadvertance ou viser délibérément à éviter la deuxième phase de collecte de données. L'erreur due aux classifications incorrectes peut produire des niveaux graves de non-couverture, particulièrement quand la classification de la population rare est fondée sur les réponses à plusieurs questions, car une réponse incorrecte à n'importe laquelle de

Dans les enquêtes où les personnes sont échantillonnées en commençant par échantillonner les ménages, les concepteurs préfèrent souvent sélectionner une personne par ménage – peut-être en permettant l'échantillonnage de deux personnes dans les grands ménages – pour éviter les effets de contamination et prévenir un effet d'homogénéité de grappes à l'intérieur des ménages sur les effets de plan. Ce plan d'échantillonnage n'est pas toujours le meilleur (Clark et Steel 2007), particulièrement quand on échantillonne des populations rares. Si les membres de la population rare sont concentrés dans certains ménages (par exemple populations minoritaires), la taille de l'échantillon de présélection risque d'être réduite sensiblement si plus d'une personne, voire même toutes les personnes admissibles, peuvent être tirées dans certains ménages (voir Hedges 1973). Selon Elliott, Finch, Klein, Ma, Do, Becker, Orr et Lurie (2008), pour le suréchantillonnage des minorités d'Amérindiens/Natifs de l'Alaska et de Chinois aux États-Unis, la sélection de toutes les personnes admissibles dans un ménage est prometteuse pour les enquêtes sur la santé américaines. Le plan d'échantillonnage de la NHANES maximise le nombre de personnes échantillonnées par ménage. Puisque chaque répondant est rémunéré pour sa participation, les ménages comptant un plus grand nombre de répondants reçoivent une plus forte rémunération, facteur que l'on pense accroître les taux de réponse (Mohadjer et Curtin 2008). Il convient de souligner que l'homogénéité intra-ménage aura peu d'incidence sur les effets de plan si les données sont analysées selon les caractéristiques de sous-groupes (par exemple l'âge et le sexe) qui recoupent tous les ménages.

Le recours à une présélection à grande échelle pour cerner les populations rares suscite trois problèmes qui peuvent chacun empêcher d'atteindre les tailles d'échantillon prévues à moins de prendre des précautions. Le premier problème découle du fait que, en cas de présélection, la taille d'échantillon d'une population rare est une variable aléatoire. Par conséquent, la taille d'échantillon obtenue peut être plus grande ou plus petite que prévu. Si une taille minimale d'échantillon est spécifiée pour une population rare, il pourrait être judicieux de déterminer la fraction d'échantillonnage qu'il convient d'utiliser pour être certain qu'il existe, disons, une probabilité de 90 % que la taille réalisée d'échantillon soit au moins aussi grande que le minimum spécifié. Cette procédure a été utilisée pour déterminer les fractions d'échantillonnage pour les nombreux sous-domaines âge-sexe-revenu pour la Continuing Survey of Food Intakes by Individuals de 1994-1996 (Goldman, Borud et Berlin 1997).

Le deuxième problème que pose la présélection à grande échelle tient au fait qu'il faut tenir compte du taux de non-réponse global. Un membre échantillonné d'une population rare sera un non-répondant si l'information du

les enfants de 19 à 35 mois en vue de déterminer les niveaux de couverture de la vaccination (Smith, Battaglia, Huggins, Hoaglin, Roden, Khare, Ezzi-Rice et Wright 2001 ; U.S. National Center for Health Statistics 2009b). La présélection à grande échelle de la NIS est également utilisée pour les membres des domaines d'intérêt du programme de la State and Local Area Integrated Telephone Survey (SLATS), qui porte sur divers autres thèmes au cours du temps (U.S. National Center for Health Statistics 2009c). Si l'on répartit les coûts de présélection entre un certain nombre d'enquêtes, il est avantageux que les domaines des enquêtes soient des ensembles assez disjoints afin de réduire au minimum de présélectionner certaines personnes pour plus d'une enquête.

Si personne n'est présent au domicile pour répondre au questionnaire de présélection administré sur place, l'information peut parfois être obtenue auprès de voisins bien informés quant à l'appartenance d'un des membres du ménage à la population rare (par exemple un enfant de moins de trois ans). Cette approche (qui est utilisée dans la NHANES) permet de réduire sensiblement les coûts de la collecte des données quand une forte proportion de ménages ne contiennent pas de membre de la population rare. Cependant, l'approche risque de produire une couverture insuffisante ; le fait d'exiger que, si le premier voisin interrogé déclare que le ménage ne comprend aucun membre de la population rare, l'autre voisin soit également interrogé offre une certaine protection. Les questions d'éthique doivent également être prises en considération, particulièrement pour l'identification des populations rares de nature sensible.

Une extension de la collecte d'information de présélection auprès des voisins porte le nom de dénombrement ciblé. Cette méthode, qui est une forme d'échantillonnage par réseaux (voir la section 3.5), consiste à demander au répondant à chaque adresse échantillonnée, ou adresse « base », si des membres de la population rare résident aux n adresses voisines de part et d'autre. Essentiellement, l'échantillon est constitué de $2n + 1$ adresses pour fournir l'information de présélection pour une ou plusieurs adresses liées, l'intervieweur doit prendre contact avec une personne résidant à une autre adresse. Le dénombrement ciblé a été utilisé avec $n = 2$ dans la British Crime Survey (Bolling, Grant et Sinclair 2008) et dans la Health Survey of England (Frens, Prior, Koroveress, Calderwood, Brookes et Primatesia 2001) pour suréchantillonner les minorités ethniques. Une limite de la méthode est qu'elle produit vraisemblablement un sous-dénombrement (éventuellement important). Des indices de l'importance du sous-dénombrement peuvent être obtenus en comparant la prévalence de la population rare dans l'échantillon de base à celle dans les adresses liées.

La taille de l'échantillon de première phase est la taille minimale d'échantillon qui produit les tailles d'échantillon requises (ou plus grandes) pour tous les domaines. La taille minimale de l'échantillon de première phase est déterminée en établissant la taille d'échantillon requise pour l'un des domaines, puis en incluant dans l'échantillon de seconde phase tous les membres de l'échantillon de ce domaine. Des sous-échantillons des autres domaines sont sélectionnés pour l'échantillon de deuxième phase à des taux qui produisent les tailles d'échantillon de domaine requises. Si l'enquête est conçue pour recueillir des données pour un seul sous-ensemble de domaines (souvent un seul domaine), aucun membre des autres domaines n'est sélectionné dans l'échantillon de deuxième phase.

Utiliser un mode peu coûteux de collecte de données, tel que l'interview téléphonique ou un questionnaire envoyé par la poste, pour la présélection. La collecte des données de deuxième phase peut se faire par le même mode ou par un mode différent.

Si cela est possible et utile, permettre la collecte de données de présélection auprès d'autres personnes que celles échantillonnées. Par exemple, d'autres membres du ménage pourraient être capables d'indiquer correctement la situation d'appartenance à la population rare du membre échantillonné. Voir la discussion plus bas, ainsi que la section 3.5 sur l'échantillonnage par réseaux.

Si la présélection se fait par interview sur place selon un plan à plusieurs degrés, choisir une grande taille d'échantillon dans chaque grappe est une mesure efficace. L'utilisation de grappes compactes est également possible. Les coûts sont réduits et la précision des estimations par domaine n'est pas gravement altérée, parce que les tailles moyennes d'échantillon de domaine dans les grappes seront relativement faibles.

Un moyen éventuel de réduire les coûts de présélection consiste à les répartir entre plus d'une enquête. Par exemple, aux États-Unis, la composante des enfants de la National Immunization Survey (NIS) est une enquête téléphonique trimestrielle destinée à présélectionner les ménages dotés d'un numéro de téléphone conventionnel à fil pour repérer

chaque domaine aient des probabilités égales de sélection peuvent être toutes deux satisfaites en échantillonnant les grappes par des méthodes PPT classique, mais en utilisant une mesure composite de la taille qui tient compte des fractions d'échantillonnage différentes pour les divers domaines (Folsom, Porter et Williams 1987). Ainsi, dans une enquête réalisée auprès des hommes dans les prisons anglaises, les fractions d'échantillonnage souhaitées étaient de 1 sur 2 pour les prisonniers civils (C), de 1 sur 21 pour les prisonniers « star » qui servent normalement leur première peine (S) et de 1 sur 45 pour les récidivistes (R). Les prisons ont été sélectionnées au premier degré d'échantillonnage, la prison i étant tirée avec probabilité proportionnelle à sa mesure composite de taille $R_i + 2,2S_i + 20,3C_i$, où les multiplicateurs sont les taux d'échantillonnage par rapport au taux pour les récidivistes (Morris 1965, pages 303 à 306).

3. Méthodes de suréchantillonnage des domaines rares

Le présent article porte principalement sur l'utilisation de méthodes d'échantillonnage probabilistes pour produire des estimations fondées sur le plan de sondage classiques, ou estimations directes, des caractéristiques de populations rares, qui sont souvent des petits domaines selon la terminologie de Kish. Avant d'aborder la discussion de ces méthodes, il serait utile de mentionner certaines caractéristiques de divers types de populations rares qui, de pair avec le mode de collecte de l'enquête, influencent le choix des méthodes d'échantillonnage applicables en vue de produire les tailles d'échantillon requises pour tous les domaines. Suit un résumé de certaines caractéristiques importantes dont il convient de tenir compte.

- Existe-t-il une ou plusieurs listes distinctes pour l'échantillonnage d'une population rare? Les personnes échantillonnées peuvent-elles être repérées pour la collecte des données? La liste est-elle à jour et complète? Si une liste à jour existante contient uniquement la population rare (avec éventuellement quelques autres listages) et qu'elle fournit une couverture presque complète, l'échantillonnage peut se faire par des méthodes classiques. Si aucune liste unique ne donne une couverture adéquate, mais qu'il existe plusieurs listes qui, entre elles, donnent une bonne couverture, des problèmes de pistes multiples de sélection se posent (section 3.4).
- La population rare est-elle concentrée dans certaines parties identifiables de la base de sondage ou y

est-elle dispersée assez uniformément? Si elle est concentrée, la stratification disproportionnée peut être efficace (section 3.2).

- Si un échantillon est sélectionné à partir d'une population plus générale, est-il possible de déterminer à peu de frais si une personne échantillonnée appartient à la population rare, par exemple d'après les réponses à quelques questions simples? Dans l'affirmative, les méthodes de présélection standard peuvent être utilisées (section 3.1). Si la détermination exacte requiert des procédures coûteuses, telles que des examens médicaux, un plan de sondage à deux phases pourrait être utile (section 3.3). Une question connexe est celle de savoir si certains membres d'une population rare considèrent que leur appartenance à cette population est une matière délicate; la probabilité que certains membres soient tentés de nier leur appartenance à la population peut influencer le choix du mode d'administration du questionnaire et d'autres aspects de la présélection.
- Les membres de la population rare peuvent-ils être identifiés facilement par d'autres? Dans l'affirmative, une certaine forme d'échantillonnage par réséaux pourrait être utile (section 3.5).
- Les membres de la population rare peuvent-ils être trouvés dans des lieux ou à des événements particuliers? Dans l'affirmative, l'échantillonnage des lieux de sociabilité pourrait être utile (section 3.6).
- La population rare est-elle définie par une caractéristique constante (par exemple race/ethnicité) ou par un événement récent (par exemple une hospitalisation)? La distinction entre ces deux types de caractéristiques est un élément important à prendre en considération pour déterminer l'utilité des enquêtes par panel pour l'échantillonnage de populations rares (section 3.7).

Une gamme de méthodes d'échantillonnage des populations rares sont passées en revue aux sections suivantes. Bien que les méthodes soient discutées individuellement, certaines sont interdépendantes et, dans la pratique, une combinaison de méthodes est souvent utilisée.

3.1 Présélection

Une certaine forme de présélection est généralement nécessaire quand la base de sondage ne contient pas d'identificateurs de domaine. La présente section décrit une application simple d'un plan de présélection dans lequel un grand échantillon de première phase est sélectionné pour définir des échantillons de membres des domaines d'intérêt, sans recourir aux techniques décrites dans les sections qui

échantillons supplémentaires pour fournir des estimations infraprovinciales d'une précision acceptable.

Les scénarios de Kish et de Bankier reposent sur l'hypothèse que le même niveau de précision doit être obtenu pour tous les petits domaines. Longford (2006) décrit une approche plus générale dans laquelle des « priorités inférieures » P_d sont assignées à chaque domaine d . Par exemple, il propose de choisir les priorités de sorte que $P_d = N_d^a$ où N_d est la taille de la population du domaine d et a est une valeur choisie entre 0 et 2. La valeur $a = 0$ correspond à l'hypothèse de taille d'échantillon de domaine égale de Kish et de Bankier, et $a = 2$ correspond à une répartition globale proportionnelle. Une valeur intermédiaire de a accorde une plus grande priorité aux grands domaines. Longford étend aussi l'approche afin d'intégrer une priorité inférieure pour l'estimation globale.

Une approche plus générale de répartition de l'échantillon est celle de la programmation mathématique qui a été proposée par un certain nombre de chercheurs (voir, par exemple, Rodríguez Vera 1982). Cette approche permet de traiter les variances de domaine inégales, les domaines qui se recoupent et les estimations multiples pour chaque domaine. L'Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) des États-Unis fournit un exemple de domaines qui se recoupent, où l'échantillon a été sélectionné d'après les enregistrements d'actes de naissance qui contenaient l'information requise sur les domaines. Pour l'ECLS-B, il existait dix domaines d'intérêt, à savoir les naissances classées selon la race (cinq domaines), le poids à la naissance (trois domaines), ainsi que les naissances jumelées ou non jumelées (deux domaines). L'approche adoptée consistait à déterminer d'abord une taille d'échantillon efficace minimale (c'est-à-dire la taille d'échantillon réelle divisée par l'effet de plan) pour chaque domaine. En traitant les 30 cellules de la classification croisée du poids à la naissance, de la race/ethnicité et de la naissance jumelée/non jumelée comme des strates, une répartition de l'échantillon entre les strates a ensuite été déterminée de manière à réduire au minimum la taille globale de l'échantillon tout en satisfaisant les exigences de taille d'échantillon efficace pour tous les domaines (Green 2000).

Lorsqu'il existe de multiples domaines d'intérêt et qu'il faut utiliser l'échantillonnage à plusieurs degrés, une variante de la mesure de taille habituelle pour l'échantillonnage avec probabilité proportionnelle à la taille (PPT) peut être utile pour contrôler les tailles d'échantillon dans les grappes échantillonnées (UPF, unités de deuxième degré, etc.), à condition que des estimations raisonnables des tailles de population de domaine soient disponibles par grappe. Les exigences que toutes les grappes échantillonnées aient approximativement la même taille globale de sous-échantillon et que les unités échantillonnées dans

les domaines sont des régions administratives du pays, telles que des États, des provinces, des comtés ou des districts. Le cas échéant, l'adoption de la répartition optimale pour un objectif donne lieu à une perte importante de précisions pour l'autre. Cependant, une répartition de compromis qui se situe entre les deux répartitions optimales donne souvent de bons résultats pour les deux objectifs.

Il existe plusieurs solutions de compromis. L'une, proposée par Kish (1976, 1988), consiste à déterminer les tailles d'échantillon de domaine en appliquant la formule

$$n_h \propto \sqrt{IW_h^2} (1 + I - I^{-2})^{-1/2}$$

où I et $(1 - I)$ représentent l'importance relative de l'estimation nationale et des estimations par domaine (par exemple district administratif), respectivement. Si $I = 1$, la répartition est proportionnelle, c'est-à-dire optimale pour l'estimation nationale, tandis que si $I = 0$, la répartition est égale, c'est-à-dire optimale pour les estimations par domaine. Le choix de I est très subjectif, mais j'ai constaté que $I = 0,5$ est souvent un bon point de départ, après lequel un examen attentif de la répartition peut entraîner des modifications. Bankier (1988) a proposé une solution de compromis similaire, dite répartition exponentielle. Si elle est appliquée à l'exemple considéré ici, les tailles d'échantillon de domaine sont déterminées d'après $n_h \propto W_h^a$, où q est une puissance comprise entre 0 (répartition égale) et 1 (répartition proportionnelle). Par exemple, l'Enquête sur la santé dans les collectivités canadiennes de 2007 a été conçue en vue d'accorder une importance à peu près égale aux estimations pour les provinces et pour les régions socio-sanitaires. La part de l'échantillon allouée à une province a été fondée sur la taille de la population de cette dernière et sur son nombre de régions socio-sanitaires. À l'intérieur d'une province, l'échantillon a été réparti entre les régions socio-sanitaires en utilisant la répartition de Bankier avec $q = 0,5$ (Statistique Canada 2008).

Une limite des méthodes de Kish et de Bankier est qu'elles peuvent ne pas allouer aux petits domaines un échantillon suffisant pour produire des estimations au niveau requis de précision. Cette limite peut être contournée en révisant les allocations initiales de manière à satisfaire les exigences de précision. Une approche alternative résout le problème que pose cette limite directement : la répartition est déterminée en fixant un échantillon de base qui satisfait l'un des objectifs, puis en complétant cet échantillon au besoin pour satisfaire l'autre objectif. Singh, Gambino et Mantel (1994) décrivent un plan d'échantillonnage de ce genre pour l'Enquête sur la population active du Canada, qui comporte un échantillon de base pour produire des estimations nationales et provinciales et, au besoin, des

convient uniquement pour les mini-domaines pour lesquels existent des réseaux bien définis. La méthode est surtout utilisée dans des conditions locales, mais Katzoff, Sirken et Thompson (2002) et Katzoff (2004) ont suggéré que les unités de départ pourraient provenir d'une enquête à grande échelle, telle que la National Health Interview Survey aux États-Unis.

Le présent article, axé sur l'utilisation des méthodes d'échantillonnage probabilistes pour produire des estimations fondées sur le plan de sondage classiques, c'est-à-dire directes, des caractéristiques des populations rares, a pour but d'offrir les revues antérieures (par exemple Kish 1965a; Kalton et Anderson 1986; Kalton 1993a, 2003; Sudman et Kalton 1986; Sudman, Sirken et Cowan 1988; Flores Cervantes et Kalton 2008). La plupart des publications traitent des problèmes d'échantillonnage qu'il faut résoudre quand la population rare est le seul sujet d'étude. Toutefois, comme il est mentionné plus haut, les enquêtes doivent souvent permettre de produire des estimations pour de nombreux domaines distincts, ainsi que pour l'ensemble de la population. La section 2 offre une revue des problèmes que pose l'élaboration du plan de sondage quand les objectifs de conception de l'enquête ont trait à des domaines multiples dont les membres peuvent être identifiés d'après les renseignements figurant dans la base de sondage. La section 3, qui est la partie principale de l'article, passe en revue diverses méthodes qui ont été utilisées pour échantillonner les populations rares dont les membres ne peuvent pas être identifiés au préalable. Enfin, la section 4 présente certaines conclusions.

2. Répartitions pour domaines multiples

La question de la répartition de l'échantillon se pose quand une enquête est conçue en vue de produire des estimations pour un certain nombre de domaines différents, pour des sous-classes qui recourent les domaines, ainsi que pour la population totale. Dans la plupart des applications, les domaines sont de tailles très variables, au moins certains étant des domaines rares.

Supposons que H domaines mutuellement exclusifs et exhaustifs sont identifiés dans la base de sondage. Sous les hypothèses habituelles que la variance d'une estimation pour le domaine h peut être exprimée par V/n_h , et que les coûts d'enquête sont les mêmes dans les divers domaines, la répartition optimale pour l'estimation de la moyenne globale de population est $n_h \propto W_h$, où W_h est la proportion de la population dans le domaine h . Si l'on suppose que les estimations par domaine doivent toutes avoir la même précision, la répartition optimale est $n_h = n/H$ pour tous les domaines. Ces deux répartitions sont en conflit quand W_h varie fortement, comme cela se produit souvent quand

données recueillies maintenant grâce à l'American Community Survey et sur des variables prédictives obtenues d'autres sources disponibles au niveau local, telles que les données fiscales (U.S. Census Bureau 2009b). Un traitement complet de l'estimation indirecte par les techniques d'estimation sur petits domaines, méthodeologie qui dépasse le cadre du présent article, peut être consulté dans Rao (2003).

À part l'échantillonnage des emplacements ou des lieux de sociabilité, dont il est question à la section 3.6, le présent article ne s'attarde pas aux diverses méthodes qui ont été élaborées pour l'échantillonnage d'autres types de mini-domaines suscitant un grand intérêt chez les chercheurs du domaine social et les épidémiologistes, domaines qui sont souvent des « populations cachées » en ce sens que les activités qui les définissent sont clandestines, comme la prise de drogues par voie intraveineuse (Watters et Biernacki 1989). Une gamme de méthodes ont été mises au point sous l'hypothèse que les membres des mini-domaines se connaissent entre eux. La classe générale des plans d'échantillonnage de ce type porte le nom de plans de sondage avec dépiçage de liens (voir la revue faite par Thompson et Frank 2000). Il s'agit de plans de sondage adaptés dans lesquels les unités sont sélectionnées séquentiellement, celles sélectionnées aux étapes ultérieures étant dépendantes de celles sélectionnées auparavant (Thompson et Seber 1996; Thompson 2002).

L'échantillonnage en boule de neige a été l'une des premières méthodes d'échantillonnage adaptées en chaîne. Il a pour point de départ un échantillon initial de membres du domaine rare (les grâces) qui, à leur tour, identifient d'autres membres du domaine. Bien qu'il ressemble à d'autres versions de l'échantillonnage en boule de neige à connues, non nulles, pour tous les membres du domaine. L'une des versions de l'échantillonnage en boule de neige a été nommée échantillonnage dirigé par les répondants (EDR) (Heckathorn 1997, 2007). Volz et Heckathorn (2008) ont élaboré une théorie du EDR qui est fondée sur les quatre hypothèses suivantes : 1) les répondants savent à combien de membres du réseau ils sont reliés (le degré), 2) les répondants en recrutent d'autre dans leur réseau personnel au hasard, 3) les connexions du réseau sont réciproques et 4) le recrutement est fait selon un processus markovien. La nécessité de formuler ces hypothèses de modélisation pour l'inférence statistique est l'élément qui distingue ces plans d'échantillonnage en chaîne des plans d'échantillonnage probabilistes classiques utilisés dans les enquêtes pour lesquelles il n'est pas nécessaire de faire ce genre d'hypothèses. Il apparaît clairement que le RDS

provinces ou les cellules de la classification croisée du groupe d'âge et de la race/ethnicité) ou se recouper (par exemple des domaines définis séparément selon le groupe d'âge et selon la race/ethnicité).

La taille d'un domaine est un élément clé dont il faut tenir compte. Kish (1987) a proposé une classification qui comprend les *grands domaines*, représentant 10 % ou plus de la population totale, pour lesquels un échantillon général produit habituellement des estimations fiables, les *petits domaines*, représentant de 1 % à 10 % de la population totale, pour lesquels les méthodes d'échantillonnage décrites dans le présent article sont nécessaires, les *mini-domaines*, représentant de 0,1 % à 1 % de la population, pour lesquels les estimations doivent être produites en utilisant principalement des modèles statistiques, et les *types rares*, comprenant moins de 0,01 % de la population, qui ne peuvent généralement pas être traités par les méthodes d'échantillonnage appliquées dans les enquêtes. De nombreuses enquêtes visent à produire des estimations pour certains grands domaines, certains petits domaines et, à l'occasion, même certains mini-domaines.

Comme les tailles d'échantillon de la plupart des enquêtes sont suffisantes pour produire des estimations d'une précision raisonnable pour les grands domaines, il n'est généralement pas nécessaire d'adapter les types de procédures de suréchantillonnage passés en revue ici. Cependant, certaines caractéristiques importantes du plan de sondage doivent être prises en considération. Ainsi, il est utile de tenir compte des grands domaines dans la création des strates de l'enquête. Cet aspect est spécialement important quand les domaines sont définis géographiquement et que l'on procède à un échantillonnage à plusieurs degrés. Si l'on ne fait pas coïncider un domaine géographique avec une strate du plan de sondage, le nombre d'unités primaires d'échantillonnage (UPF) sélectionnées dans ce domaine est une variable aléatoire ; les UPF échantillonnées dans les strates qui couvrent les limites du domaine peuvent ou non être dans le domaine, ce qui pose des problèmes pour l'estimation par domaine. Il est également intéressant d'obtenir un nombre appréciable d'UPF échantillonnées dans chaque domaine géographique pour pouvoir calculer des estimations de variance directes d'une précision raisonnable, ce qui implique qu'il faut étaler l'échantillon sur un grand nombre d'UPF. À l'étape de l'estimation, il est préférable, si cela est possible, d'appliquer les corrections de la non-réponse et de la non-couverture de type poststratification au niveau du domaine plutôt qu'au niveau national. Singh, Gambino et Mantel (1994), ainsi que Marker (2001) discutent des questions concernant le plan de sondage et Rao (2003, pages 9 à 25) s'intéresse aux problèmes d'estimation pour les grands domaines. Peu d'attention sera accordée à ces derniers dans le présent article.

À l'autre extrémité du spectre de tailles, même si l'on applique des méthodes d'échantillonnage probabilistes spéciales, les tailles d'échantillon possibles pour la plupart des enquêtes ne sont pas suffisamment grandes pour produire des estimations classiques fondées sur le plan de sondage, ou estimations directes, des caractéristiques pour de multiples domaines quand un grand nombre d'entre eux sont des mini-domaines ou des types rares. Un recensement national de population est une exception manifeste, mais même les recensements ont leurs limites. Puisqu'ils sont réalisés peu fréquemment (dans de nombreux pays une fois tous les dix ans seulement), les estimations qui en résultent sont dépassées, ce qui est particulièrement préoccupant pour les mini-domaines, dans lesquels les changements peuvent être rapides. En outre, le contenu d'un recensement est fortement limité en ce qui concerne la gamme des sujets abordés et la profondeur du détail. De très grandes enquêtes continues, telle l'American Community Survey (U.S. Census Bureau 2009a; Citro et Kalton 2007), le recensement continu de la France (Durr 2005) et le microrecensement d'Allemagne (German Federal Statistical Office 2009) ont été élaborés pour répondre aux besoins de données plus à jour pour les petits domaines, mais une restriction quant au contenu persiste (quoique le contenu du microrecensement allemand varie au cours du temps). D'autres exceptions ont lieu à la limite entre les mini-domaines et les petits domaines. Par exemple, depuis 2007, l'Enquête sur la santé dans les collectivités canadiennes fournit des estimations sur l'état de santé des populations de chacune des 121 régions sociosanitaires du Canada basées sur une enquête annuelle réalisée auprès d'environ 65 000 personnes de 12 ans et plus avec production de fichiers de données annuels et bis-annuels (Statistique Canada 2008). En combinant les échantillons de plusieurs années, les chercheurs peuvent produire des estimations pour divers types de populations rares.

Toutefois, en général, la taille maximale d'échantillon possible pour une enquête portant sur un sujet particulier ne convient pas pour produire un grand ensemble d'estimations par mini-domaine d'une précision acceptable. Pourtant, les décideurs continuent d'accroître la demande de données locales au niveau du mini-domaine. Pour répondre à cette demande d'estimation pour des mini-domaines, qui sont principalement des domaines définis au moins en partie par des unités administratives géographiques, on recourt à des techniques de modélisation statistiques qui mènent à des estimations sur petits domaines indirectes, dépendantes d'un modèle. Ainsi, le programme des Small Area Income and Poverty Estimates du U.S. Census Bureau produit chaque année des estimations indirectes du revenu et des statistiques sur la pauvreté pour 3 141 comités et des estimations des enfants d'âge scolaire vivant dans la pauvreté pour environ 15 000 districts scolaires, en se fondant sur des

Méthodes de surechantillonnage des sous-populations rares dans les enquêtes sociales

Graham Kalton¹

Résumé

Souvent, les enquêtes doivent permettre de produire des estimations pour une ou plusieurs sous-populations en plus de l'ensemble de la population. Lorsque l'appartenance à une sous-population (ou domaine) rare peut être déterminée d'après l'information contenue dans la base de sondage, le choix de la taille de l'échantillon du domaine est relativement simple. Le principal problème consiste alors à déterminer l'ampleur requise du surechantillonnage quand des estimations doivent être produites pour plusieurs domaines ainsi que pour l'ensemble de la population. En revanche, l'échantillonnage et le surechantillonnage de domaines rares dont les membres ne peuvent pas être identifiés d'avance posent un défi important. Diverses méthodes ont été utilisées dans cette situation. En plus de la présélection à grande échelle, elles comptent l'échantillonnage stratifié disproportionné, l'échantillonnage à deux phases, l'utilisation de plusieurs bases de sondage, l'échantillonnage par réseaux, l'échantillonnage des lieux de sociabilité, les enquêtes par panel et les enquêtes polyvalentes. Le présent article décrit l'application de ces méthodes à une gamme d'enquêtes sociales.

Mots clés : Répartition de l'échantillon ; présélection ; échantillonnage stratifié disproportionné ; échantillonnage à deux phases ; bases de sondage multiples ; échantillonnage des lieux de sociabilité ; enquêtes par panel ; enquêtes polyvalentes.

1. Introduction

C'est pour moi un très grand privilège d'avoir été choisi comme auteur de l'article sollicité de la série Waksberg, qui met Joe Waksberg à l'honneur pour ses nombreuses contributions aux techniques d'enquête. J'ai eu la chance de travailler avec Joe à Westat pendant de nombreuses années et, comme pour beaucoup d'autres, l'expérience a été enrichissante. Quand il était confronté à un problème d'échantillonnage insoluble, Joe n'avait pas son pareil pour le retourner et trouver une solution pratique. Puisque le problème concernait souvent l'échantillonnage de populations rares, j'ai choisi pour thème du présent article les méthodes d'échantillonnage de ces populations.

Au cours des dernières décennies, l'un des principaux faits nouveaux dans le domaine des études par sondage a été la croissance continue de la demande d'estimations pour des sous-classes (sous-populations) de plus en plus petites de la population générale. Le présent article porte sur les sous-classes – appelées *domaines* – dont on prévoit l'analyse distincte au moment de l'élaboration du plan de sondage. Les États ou provinces, les comités ou les districts d'un pays, les minorités raciales ou ethniques, les ménages vivant dans la pauvreté, les naissances récentes, les personnes de plus de 80 ans, les nouveaux immigrants, les homosexuels, les toxicomanes et les personnes handicapées sont des exemples de domaines qui ont été pris en considération dans l'élaboration des plans de sondage de diverses enquêtes. Quand les domaines sont petits (on parle aussi de *populations rares*), la nécessité d'obtenir des tailles d'échantillon

adéquates pour l'analyse par domaine crée parfois de grands défis en ce qui concerne le plan de sondage. Le présent article offre une revue des diverses méthodes d'échantillonnage probabilités qui ont été utilisées pour générer des échantillons permettant d'estimer les caractéristiques de populations rares avec le niveau requis de précision. Les méthodes d'échantillonnage permettant d'estimer la taille d'une population rare n'y sont pas abordées explicitement, quoique des méthodes comparables aux précédentes soient souvent applicables. Toutefois, le présent article ne traite pas des méthodes de capture-recapture ni des méthodes connexes.

Une question importante à prendre en considération lors de l'élaboration d'un plan de sondage est celle de savoir si l'enquête est destinée à produire des estimations pour un seul ou pour de nombreux domaines. Alors que la plupart des publications sur l'échantillonnage de populations rares portent sur les plans d'échantillonnage pour un seul domaine rare (par exemple, les toxicomanes), en pratique, les enquêtes sont souvent conçues en vue de produire des estimations pour de nombreux domaines (par exemple, chaque province d'un pays ou plusieurs groupes raciaux/ethniques). La National Health and Nutrition Examination Survey (NHANES) réalisée aux États-Unis est un exemple d'enquête conçue pour produire des estimations pour de nombreux domaines, dans ce cas, définis selon l'âge, le sexe, la race/l'ethnicité et la situation de faible revenu (Mohadjer et Curtin 2008). Dans les plans d'échantillonnage qui comprennent de nombreux domaines, ces derniers peuvent être mutuellement exclusifs (par exemple, les

Membres du comité de sélection de l'article Waskberg (2009-2010)

Daniel Kasprzyk (Chair), *Mathematica Policy Research*
Wayne A. Fuller, *Iowa State University*
Elizabeth A. Martin
Mary Thompson, *University of Waterloo*

Présidents précédents :

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007-2008)
Leyla Mojadjer (2008-2009)

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg. L'auteur reçoit une prime en argent qui provient d'une bourse de Westat, en reconnaissance des contributions de Joe Waksberg pendant ses nombreuses années de collaboration avec Westat. L'administration financière de la bourse est assurée par l'American Statistical Association.

Gagnants du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Sæmndal (2007)
Mary Thompson (2008)
Graham Kalton (2009)
Ivan Fellegi (2010)

Nominations :

L'auteur de l'article Waksberg de 2011 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées à Daniel Kasprzyk, présidente du comité, par courriel à DKasprzyk@Mathematica-Mpr.com. Les candidatures et les suggestions de sujets doivent être reçues d'ici au 28 février 2010.

Article sollicité Waksberg 2009

Auteur : Graham Kalton

Graham Kalton est président du comité de direction et vice-président directeur de Westat. Il détient le titre de professeur de la recherche du Programme conjoint de la méthodologie d'enquête de l'Université de Maryland. Ayant des intérêts très variés dans la méthodologie d'enquête, Dr Kalton a publié des articles portant sur plusieurs facettes de la méthodologie, y compris le plan de sondage, la non-réponse et l'imputation, les enquêtes par panel, la formulation de questions et le codage. Il est membre de l'American Association for the Advancement of Science, de l'American Statistical Association et du National Association of the National Academies et est dirigeant élu de l'International Statistical Institute. Il a, en outre, prononcé le discours annuel Morris Hanson en 2000.

Shao et Thompson étudient le problème de l'estimation de la variance lorsque l'on procède à un ajustement de la pondération pour corriger la non-réponse dans les enquêtes-entreprises stratifiées. Ils deviennent deux estimateurs de variance jackknife naïfs ne donnant pas de bons résultats à moins que la fraction d'échantillonnage soit négligeable, ce qui n'est pas le cas en présence de strates à tirage complet. Ils proposent un estimateur de variance jackknife modifié et convergent même en présence de strates à tirage complet, mais la fraction d'échantillonnage des strates à tirage partiel ne doit pas être élevée. Ils évaluent leur estimateur de variance empiriquement au moyen de données réelles et par une étude en simulation.

Dans son article, Preston étudie l'estimation de variance bootstrap pour les plans à plusieurs degrés quand les unités sont sélectionnées par échantillonnage aléatoire simple sans remise à chaque degré. Il propose une extension de l'estimateur bootstrap rééchantillonné souvent utilisé, qui repose sur l'hypothèse d'un échantillonnage avec remise ou de fractions d'échantillonnage négligeables au premier degré. Il compare l'estimateur proposé aux estimateurs bootstrap rééchantillonné et bernoillien. Jiang et Eltinge s'attaquent au problème de l'estimation du nombre de degrés de liberté sous des plans de sondage multivariés à plusieurs degrés quand un petit nombre d'unités primaires d'échantillonnage (UPF) sont tirées par strate. Étant donné le petit nombre d'UPF sélectionnées, l'estimation classique du nombre de degrés de liberté selon la méthode de Satterthwaite peut donner lieu à une sous-estimation grave. Les auteurs proposent un estimateur de rechange du nombre de degrés de liberté qui utilise les variances intra-UPF pour fournir de l'information auxiliaire sur la grandeur relative de la variance globale au niveau de la strate. Ils illustrent leur méthode au moyen de données provenant de la National Health and Nutrition Examination Survey (NHANES).

L'article de Wang et Bellhouse explore l'application de méthodes de régression non paramétriques pour étudier la relation entre la variable réponse et les covariables, ainsi que la prédiction en utilisant de l'information auxiliaire dans le contexte des enquêtes complexes. Leurs travaux prolongent ceux de Bellhouse et Stafford (2001), qui ont appliqué une fonction de régression non paramétrique simple dans l'analyse par régression portant sur des données d'enquête.

Et finalement, nous sommes heureux d'informer les lecteurs et auteurs que *Techniques d'enquête* sera bientôt citée par SCOPUS sur les bases de données Elsevier Bibliographic Databases, et ce à partir du numéro de juin 2008.

Harold Mantel, rédacteur en chef délégué

Dans ce numéro

Ce numéro de *Techniques d'enquête* commence par le neuvième article de la série annuelle à la théorie et à la pratique des techniques d'enquête. Le comité de rédaction remercie les membres Wayne Fuller, d'avoir choisi Graham Kalton comme auteur de l'article du prix Waksberg de cette année.

Dans son article intitulé « Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales », Kalton donne un aperçu des méthodes d'échantillonnage de populations rares, que Kish avait appelées « domaines mineurs ». Après avoir abordé des questions générales, il décrit un certain nombre de méthodes, dont le dépistage, la stratification, l'échantillonnage à deux phases, les bases de sondage multiples, l'échantillonnage basé sur la multiplicité, l'échantillonnage d'emplacements ou des lieux de sociabilité et l'accumulation d'échantillons au fil du temps. Il discute des avantages et des inconvénients de chacune et donne de nombreux exemples de leur utilisation dans des enquêtes. En pratique, on a souvent recours à une combinaison de diverses approches.

Des stratégies de réponse aléatoire sont souvent utilisées afin de réduire les erreurs non dues à l'échantillonnage, comme la non-réponse et les erreurs de mesure. De telles stratégies peuvent aussi être utilisées dans le contexte du contrôle de la divulgation statistique dans les fichiers de microdonnées à grande diffusion. Dans son article, Quatember propose une standardisation des méthodes de réponse aléatoire. Il établit les propriétés statistiques de l'estimateur standardisé et applique la méthode proposée à une enquête sur la tricherie chez les étudiants.

Xu et Lavallée examinent le problème causé par la non-réponse de lien dans l'échantillonnage indirect lorsqu'on utilise la méthode généralisée du partage des poids. L'échantillonnage indirect consiste à sélectionner des échantillons dans une population qui n'est pas la population cible, mais qui y est reliée. Des estimations biaisées peuvent être obtenues quand on ne sait pas qu'une unité de la population échantillonnée est liée à une unité de la population cible. Les auteurs proposent plusieurs ajustements de la pondération pour surmonter le problème de la non-réponse de lien. Dans le contexte des cas de non-réponse, les pondérations des répondants sont souvent ajustées par l'inverse de la probabilité de réponse estimée. Da Silva et Opsomer proposent de recourir à la régression par polynômes locaux pour estimer les probabilités de réponse. Ils présentent les résultats d'une étude par simulation qui confirment l'efficacité de la méthode.

L'article de Van den Brakel et Krueg porte sur un modèle de séries chronologiques structurel multivarié qui tient compte du plan de sondage de l'enquête sur la population active des Pays-Bas. Ils proposent accroit considérablement la précision des estimations.

Zhang étudie la production d'estimations croisées où l'une des marges du tableau croisé correspond à des petits domaines et où la non-réponse varie d'un domaine à l'autre. Il élabore une approche de modélisation mixte double qui combine les effets fixes et les effets aléatoires de création du modèle d'estimation sur petits domaines avec les effets aléatoires du mécanisme de conditionnelle de prédiction sous la forme d'une décomposition en trois parties qui correspondent à une variance de prédiction naïve, une correction positive qui tient compte de l'incertitude hypothétique d'estimation des paramètres fondée sur les données complètes latentes et une autre correction positive pour la variance supplémentaire due aux données manquantes.

Souza, Moura et Mignon proposent une application bayésienne d'estimation sur petits domaines fondée sur des modèles de croissance qui tiennent compte des relations hiérarchiques et spatiales. Il se sert de cette approche pour obtenir des prédictions de population pour les municipalités non échantillonnées dans l'enquête annuelle sur les ménages du Brésil et pour accroître la précision des estimations basées sur le plan de sondage obtenues pour les municipalités échantillonnées.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 35, numéro 2, décembre 2009

Table des matières

Dans ce numéro.....	129
Article Sollicité Waksberg	
Graham Kalton	
Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales.....	133
Articles Réguliers	
Andreas Quatember	
Une standardisation des stratégies fondées sur la réponse aléatoire.....	153
Xiaojian Xu et Pierre Lavallée	
Traitements de la non-réponse de lien dans l'échantillonnage indirect.....	165
Damião N. da Silva et Jean D. Opsomer	
Pondération par la propension à répondre non paramétrique fondée sur la régression par polynômes locaux pour corriger la non-réponse aux enquêtes.....	179
Jan van den Brakel et Sabine Krieg	
Estimation du taux de chômage mensuel par modélisation structurelle de séries chronologiques dans un plan de sondage avec renouvellement de panel.....	193
Li-Chun Zhang	
Estimation de la composition sur petits domaines en présence de données manquantes informatives.....	209
Debora F. Souza, Fernando A.S. Moura et Helio S. Migon	
Prediction de la population de petits domaines au moyen de modèles hiérarchiques.....	221
Jun Shao et Katherine J. Thompson	
Estimation de la variance en présence de non-répondants et de strates à tirage complet.....	235
John Preston	
Bootstrap rééchantillonné pour l'échantillonnage stratifié à plusieurs degrés.....	247
Donsig Jang et John L. Eltinge	
Utilisation des variances à l'intérieur des unités primaires d'échantillonnage pour évaluer la stabilité d'un estimateur classique de variance fondé sur le plan de sondage.....	255
Zilin Wang et David R. Bellhouse	
Modèle de régression semiparamétrique pour les données d'enquêtes complexes.....	267
Remerciements.....	283

Techniques d'enquête est répertoriée dans The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

J. Kovar

Anciens présidents

D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platak (1975-1986)

S. Fortier (Gestionnaire de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

J. Kovar, *Statistique Canada*

chef délégué

H. Mantel, *Statistique Canada*

Rédacteurs associés

J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Judkins, *Westat Inc.*
D. Kasprzys, *Mathematica Policy Research*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistique Canada*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*
D. Pfeiffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillie, *Université de Neuchâtel*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découplant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études de démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, (re@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (65 × 2 exemplaires), autres pays, 20 \$ CA (105 × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale des Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.



Ottawa

ISSN 0714-0045

Périodicité : semestrielle

N° 12-001-XPB au catalogue

Décembre 2009

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2009

Publication autorisée par le ministre responsable de Statistique Canada

Décembre 2009 • Volume 35 • Numéro 2

Une revue éditée par Statistique Canada

Techniques d'enquête

13673

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements
Service national d'appareils de télécommunications pour les malentendants
Télécopieur
1-800-263-1136
1-800-363-7629
1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements
Télécopieur
1-613-951-8116
1-613-951-0581

Programme des services de dépôt

Service de renseignements
Télécopieur pour le Programme des services de dépôt
1-800-635-7943
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de choisir la rubrique « Publications ».

Ce produit n° 12-001-X au catalogue est aussi disponible en version imprimée standard au prix de 30 \$CAN. L'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste
- Finances
Immuable R.-H.-Coats, 6^e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

Techniques d'enquête

N° 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Décembre 2009

•

Volume 35

•

Numéro 2



Statistique
Canada
Statistics
Canada

Canada

